DOCUMENT RESUME

ED 367 686 TM 021 136

AUTHOR Wheeler, Patricia; Haertel, Geneva D.
TITLE Resource Handbook on Performance Assessment and Measurement: A Tool for Students, Practitioners, and Policymakers.
REPORT NO ISBN-1-884690-15-7
PUB DATE 93
NOTE 222p.
AVAILABLE FROM Owl Press, P.O. Box 89, Berkeley, CA 94701 ($18.95, prepaid; discount available on 10 or more copies; California residents include 8.25% sales tax). available on 10 or more copies).
PUB TYPE Reference Materials - Vocabularies/Classifications/Dictionaries (134) -- Reference Materials - Bibliographies (131) -- Reference Materials - Directories/Catalogs (132)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Definitions; *Educational Assessment; *Glossaries; Measurement Techniques; Policy Formation; Reference Materials; *Resource Materials; *Resources; Standards; Test Theory
IDENTIFIERS *Alternative Assessment; *Performance Based Evaluation

ABSTRACT

This handbook addresses the complex and expanding vocabulary of performance assessment and measurement by providing a glossary of related terms and lists of resources for the student, practitioner, and policymaker. Performance assessment includes all forms of such assessment from multiple choice and paper-and-pencil tests to alternative assessments. Glossary terms were selected by rating the frequency of use and importance in explaining key concepts. Sources of additional information include professional associations and agencies, major test publishers, books and other documents on measurement and assessment, and information on professional and technical standards for developing and using assessment instruments and programs. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY
PATICIA WHEELER
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED 367 686

Resource Handbook on Performance Assessment and Measurement:

A Tool for Students, Practitioners, and Policymakers

TM021134

Patricia Wheeler, M.B.A., Ph. D.

Geneva D. Haertel, Ph.D.

BEST COPY AVAILABLE

ERRATUM:

p. 163

National Council on Measurement in Education
1230 Seventeenth Street, N.W.
Washington, DC 20036

Resource Handbook on Performance Assessment and Measurement:

A Tool for Students, Practitioners, and Policymakers

Copyright © 1993 by Patricia Wheeler and Geneva D. Haertel.

All rights reserved. Printed in the United States of America. No part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written permission of the authors.

For information, write to:

EREAPA Associates
2840 Waverley Way
Livermore, CA 94550-1740

The Owl Press
P. O. Box 89
Berkeley, CA 94701

ISBN 1-884690-15-7

First Printing

CONTENTS

Preface *vii*

Introduction 1

Glossary 13

Resources in Performance Assessment and Measurement:

- Associations and Agencies 159
- Major Test Publishers 165
- Printed Materials:
  - Books, Chapters 173
  - Articles 186
  - NCME Instructional Modules 193
  - Sources of Information on Tests 195
  - Other Reference Books 197

Standards for Assessment Programs and Tests 199

References 207

About the Authors 215

PREFACE

This *Resource Handbook* addresses the increasingly complex and expanding vocabulary of performance assessment and measurement. We were alerted to the need for a glossary on this topic by the misuse of terms in conference presentations, articles, workshops, and conversations with other professionals. The glossary section of this *Resource Handbook* lists over 500 terms that are frequently used and provides brief explanations of the meanings of these terms.

Performance assessment, as presented in this book, covers all forms of such assessment, including multiple-choice and paper-and-pencil tests as well as alternative assessments, which have been used increasingly in testing programs in recent years. Although many of the definitions and examples come from the field of education, they are applicable to human resources management in a variety of careers, training areas, and work settings, and to special programs that need to assess participant behavior.

The explanations of the terms in the "Glossary" are based upon the references listed in the back of this book, as well as the professional experiences of the authors with input and suggestions from the reviewers and several experts. Responsibility for the content of the definitions provided in the glossary section of this *Resource Handbook* rests with the authors, who, after numerous discussions, concurred on the definitions as provided.

The entries in the glossary section include new terms as well as ones that are well established. The terms were selected for frequency of use in the field, as well as

their importance in understanding key concepts related to performance assessment and measurement. Since this section of the *Resource Handbook* is a glossary and not a thesaurus or encyclopedia, the explanations are brief and have been written using neutral language to minimize any appearance of bias toward a particular approach or perspective. Many of the terms are clarified through the use of examples. Readers should be cautious, however, about applying these examples in particular settings where they may not be appropriate.

Extensive cross-referencing has been done to help the reader better understand the terms and how they differ from other terms that sometimes are used in their place. When looking up a term, the reader is encouraged to look at all the terms listed in the brackets at the end of the entry for that term. Sometimes the cross-referenced terms will contain illustrations of the original term (for example, "correlation" has some illustrations for "scatterplot"). If multiple definitions are provided for a term, they are numbered in bold type.

Sources of additional information on performance assessment and measurement are listed at the end of this volume. They include professional associations and agencies that provide materials and services on performance assessment and measurement, major test publishers in the United States, books and other documents on measurement and performance assessment, and information on professional and technical standards for developing and using assessment instruments and programs.

It is hoped that this *Resource Handbook* will lead to better communication and improved assessment and measurement practices through shared understanding of

key terms and concepts and familiarity with resources by the various individuals involved.

The following people assisted us with the preparation of and reviewed drafts of this *Resource Handbook*:

Rasjidah Franklin, Associate Professor (on leave) Dominican College, San Rafael, California, and Visiting Scholar in Education, University of California at Berkeley

John Fremer, Senior Development Leader, College Board Programs, Educational Testing Service, Princeton, New Jersey

Evelyn Hawkins, Policy Analyst, SRI International, Menlo Park, California

Dena Imlay, Training Consultant, Baton Rouge, Louisiana

Jerry A. Jenkins, Director, Region C Chapter 1 Technical Assistance Center, Educational Testing Service, Tucker, Georgia

John McNees, Instructor, Trades and Industries, San Francisco Community College, San Francisco, California, and Quality of Work Line Facilitator, U.S. Postal Service, San Francisco, California

William Mehrens, Professor of Educational Measurement, Michigan State University, East Lansing, Michigan

Gale H. Roid, Associate Professor of Psychology, George Fox College, Newberg, Oregon, and Statewide Assessment Consultant, Oregon Department of Education, Salem, Oregon

Darlene Russ-Eft, Director of Research, Zenger-Miller, San Jose, California

Joanne Walen, English Program Coordinator, Washoe County School District, Reno, Nevada

The reviewers provided us with varying perspectives from several fields of expertise and from across the United States. We benefited greatly from the comments and wisdom they generously shared with us and their willingness to help us as we refined definitions and added terms. They helped us improve the quality, accuracy, and clarity of the materials in this book. In spite of their many contributions, the authors take ultimate responsibility for the contents of this *Resource Handbook*.

Special appreciation also goes to Jean Martinson for editing several earlier versions of the *Resource Handbook*, to Jeanne Ewing of The Owl Press for encouraging us and guiding us through the publication process, to Michael Burton and Elizabeth Kilkenny Wheeler for doing the illustrations and helping the authors through their computer challenges, and to Paul K. Wheeler for helping to put the finishing touches on the manuscript.

The authors plan to publish a second edition of this *Resource Handbook*. We welcome suggestions for changes and additions for the next edition. They may be sent to Dr. Patricia Wheeler, EREAPA Associates, 2840 Waverley Way, Livermore, CA 94550-1740.

Patricia Wheeler, M.B.A., Ph.D.
Geneva D. Haertel, Ph.D.

INTRODUCTION

Performance assessment and measurement have played important roles for thousands of years. However, during the past 50 years, their prominence in our schools and workplaces has increased greatly. We use the terms performance assessment and measurement daily without having a clear understanding of what they mean or how they differ.

Performance assessment is the process of estimating the value of or describing the performance attributes of or products created by an individual or a group. Measurement is the process of classifying or estimating, in relation to a scale, rubric, or standard, the quality or quantity of an aspect of behavior, learning, or performance. Assessment is broader than measurement. We can assess how well a pianist played (i.e., make an overall judgment about the performance), or we can count the number of mistakes the pianist made (one measure of the performance's quality). The first is assessment, but not measurement since it was not quantified or linked to a scale or rubric. The second is both assessment and measurement. Confused about such terms? So are many others involved in assessment and measurement. These include people in the following arenas, among others:

> Business and industry - employees, supervisors, managers, boards of directors, union officials, professional associations, media
>
> Education - students, teachers, parents, supervisors, administrators, school board members, researchers, professional associations

Government and military - employees, enlisted personnel, supervisors, officials, military officers, legislators, judges, executive agencies, licensing boards

Health and social programs - clients, service providers, administrators, directors, policy boards

Unfortunately, confusion and miscommunication concerning these terms tend to seriously compromise the value of assessment for improving performance of individuals and groups and, ultimately, the quality of products and services. The authors believe, therefore, it is important that individuals concerned with performance assessment and measurement in all fields use terminology accurately and consistently.

We became aware of the need to improve understanding of terminology when we saw different terms being used interchangeably: assessment and evaluation, criteria and standards, benchmarks and goals, alternative assessment and authentic assessment, for example. Dictionaries and thesauri are available for evaluation, statistics, methodology, and other related areas, but none focus on assessment and measurement. We decided to address this need by developing a glossary of terms. Definitions in a glossary often do not meet all the needs of the reader, so we expanded our original effort to include information on resources. We also have included lists of associations and agencies, test publishers, and printed materials so readers can obtain additional information on topics in the glossary section. Although the lists provided are not intended to be all encompassing, we have made an effort to identify key items (e.g., landmark writings, comprehensive treatments of critical topics) and to include items for both lay audiences and experts. We have included resources from education as well as business and industry, so that

readers might learn from practitioners and researchers outside their current specialty.

Historical Context

Performance assessment has been of increasing interest in many sectors of society in recent years. But it has been in existence for many centuries. In prehistoric times, adolescent males almost certainly underwent some type of assessment before being sent out with hunting parties or on their own to bring back meat for the tribe. Historians indicate that the Chinese used some type of civil service testing 4000 years ago (DuBois, 1966). They say that, in 2200 B.C., the Emperor of China assessed the performance of officials every three years and, based on the results, either dismissed or promoted them. Formal athletic competitions were held in the Stadium of Olympia in Greece starting in 776 B.C.

During the second half of the 19th century and the early 20th century, technical methodologies for assessment were established. In Germany, early experimental psychologists studied human perception by measuring reaction time, auditory discrimination, and other aspects of sensory performance. Following the popularization of Charles Darwin's theories, individual differences and mental traits became topics of interest in Great Britain. By examining the mathematics examination scores of college students at Cambridge and admissions examination results for the Royal Military College, Sir Francis Galton found that the distributions of such scores tended to approximate a normal curve. He went on to develop covariance analysis for studying individual performance. Later, Karl Pearson and Charles Spearman developed statistical methods still in use

today in measurement and data analysis. (Crocker and Algina, 1986)

Based on the developments in Germany and Great Britain and on the burgeoning interest in human intelligence, Frenchmen Alfred Binet and Theophile Simon developed a technique for measuring human intelligence. They produced the first scale of age norms, or mental ages. They moved testing from the research laboratory to schools, medical clinics, military organizations, and workplace settings. Over the next 35 years, several versions of their intelligence tests were developed and used throughout the United States.

A Brief Chronology of Assessment and Measurement in the United States

Several key developments in performance assessment and measurement occurred in the United States in the 19th century. In 1845, Horace Mann set up a uniform written examination for Boston Public Schools (same questions given to all students). This practice spread to other school districts and led to the development in 1865 of the New York State Regents Examinations program. This program administers uniform tests in key subject areas twice a year to high school students throughout New York State. Such a program has had an impact on what content is covered by the teachers (Cronbach, 1960).

In the late 1890s, G. Stanley Hall, the founder of the child study movement, promoted the use of questionnaires to measure behaviors, attitudes, and interests of both children and adults.

Licensure programs for medical professions began in the mid-1800s. The first non-medical professional licensing examination required by a state was for barbers in Oregon in 1899 (Shimberg, 1982). Otherwise, professions and occupations tended to be self-regulating in 19th century America.

The impetus for using standardized achievement tests in America's classrooms came in 1904 with the publishing of the first textbook on testing by Professor E. L. Thorndike of Teachers College, Columbia University. Titled *An Introduction to the Theory of Mental and Social Measurement*, it helped testing to become a field of study in higher education, i.e., a specialty (Crocker and Algina, 1986). The test publishing industry started, though it was small by today's standards. Examples of tests on school subjects published in the decade following the publication of Thorndike's book are: Stone's Arithmetic Reasoning Test (1908), Courtis' Arithmetic Computation Test (1909), Thorndike's Scale for Measuring the Handwriting of Children (1910), Ayres' Scale for Measuring the Quality of Handwriting of School Children (1911), and Hillegas' Composition Scale (1912).

During World War I, the U.S. Army utilized the Army Alpha and Beta Tests (verbal and nonverbal measures of mental ability) and the Woodworth Personal Data Sheet (a personality adjustment inventory) in the selection of individuals for the military.

In the 1920s and 1930s, testing became a full-fledged and growing industry in the United States. Many articles were written on testing. Researchers used tests in their studies. More textbooks on testing were published and testing courses were offered in universities.

The first test battery--the Stanford Achievement Tests--was published in 1923. Three more achievement batteries appeared in the 1930's--the California Achievement Tests, the Iowa Tests of Basic Skills, and the Metropolitan Achievement Tests (Rudman, 1987). The first standardized interest inventories--the Strong Vocational Interest Blank (1927) and the Kuder Preference Record (1934)--were published.

The first College Board Scholastic Aptitude Test was administered to 8,000 college applicants in 1926. Law schools began using the Ferson-Stoddard Law Aptitude Test for student selection in 1927. The Association of American Medical Colleges began assessing applicants with the Moss Medical Aptitude Test in 1930. The Moss Test for Ability to Sell (1929) was being used to select salesmen and the National Teachers Examinations (1939) began to be used by school districts for the hiring of teachers.

The series of publications by Oscar Krisen Buros started with the issuance of *Educational, Psychological, and Personality Tests of 1933 and 1934* and, in 1938, the first in a series of *Mental Measurements Yearbooks* (now up to eleven). In 1939, Gertrude Hildreth issued *A Bibliography of Mental Tests and Rating Scales* (Anastasi, 1954). It covered 50 years and had over 4,000 entries. A supplement, issued in 1945, added about 1,000 more entries.

During World War II, testing become a specialized profession. The U.S. military made extensive use of tests such as the Army General Classification Test for selection and placement. The United States Armed Forces Institute (USAFI) established a program by which veterans could earn course credit and obtain a

high school diploma. This effort led to the Tests of General Educational Development (the GED Tests).

The first issue of the journal, *Psychometrika*, was published in 1936. The journal, *Educational and Psychological Measurement*, began in 1941, and, a few years later, *Personnel Psychology* was first published. Professional schools, such as nursing and teaching, started to use tests to help select candidates. Following World War II, testing specialists in the military worked in the civilian testing arena, adapting military tests to civilian uses. These led to the U.S. Employment Service's General Aptitude Test Battery and The Psychological Corporation's Differential Aptitude Test.

The use of tests by colleges and by graduate and professional schools grew as the need for selection and placement of veterans expanded rapidly at the end of World War II. In 1947, Educational Testing Service was founded to administer these large and growing testing programs for the American Council on Education, the College Entrance Examination Board (the Admissions Testing Program), and the Carnegie Foundation for the Advancement of Teaching (the Graduate Records Examinations).

In the 1950s, three more achievement batteries for schools were added to the four published prior to World War II--the Comprehensive Tests of Basic Skills, the Sequential Tests of Educational Progress, and the SRA Achievement Tests (Rudman, 1987). The American College Testing (ACT) Program was established in 1959.

Concerns about having high standards for the development and use of tests were addressed by four professional associations. In 1954, the American Psy-

chological Association (APA) published *Technical Recommendations for Psychological Tests and Diagnostic Techniques*. The next year, the National Education Association (NEA) published *Technical Recommendations for Achievement Tests*, a document developed by a joint committee of the American Educational Research Association (AERA) and the National Council on Measurement Used in Education. A third document, replacing these two, was prepared by a joint committee of APA, AERA, and the National Council on Measurement in Education (NCME), and was published in 1966 by APA. These 1966 standards have undergone two revisions, the most recent ones issued in 1985.

The testing industry continued to grow in the 1960s and 1970s as tests were used increasingly by schools to measure student achievement, by industry and business to select employees, by colleges and graduate and professional schools to select students, and by licensing agencies. The passage of the Elementary and Secondary Education Act by Congress in 1965 established and funded several categorical programs for needy students. These programs required testing of students as part of their annual evaluations. The National Assessment of Educational Progress (NAEP) was initiated in 1964 and started testing in 1969 to keep Congress and the public informed about the achievement of America's young students. Accountability became a key concern, and there was a focus on the assessment of basic skills (reading, writing, math) and on minimum competencies. Employers were concerned about the quality of candidates being turned out by the public schools.

The 1971 ruling of the U.S. Supreme Court in the *Griggs v. Duke Power Company* case focused attention

on the needs for test validation and for fairness of assessments. The next year, the Equal Employment Opportunity Act, which amended the 1964 Civil Rights Act, was passed, and in 1973 the Equal Employment Opportunity Coordinating Council issued the draft version of the now widely known *EEOC Uniform Guidelines on Employee Selection Procedures* (1978). Although the *EEOC Guidelines* applied only to employment, the importance of test validation was supported by the *Debra P. v. Turlington* (1979-84) case, in which the Florida Department of Education was required to show that the functional literacy test needed for high school graduation was a fair measure of what the schools had taught their students.

By the latter half of the seventies, the focus changed from basic skills and minimum competency to excellence, with a focus on high expectations and standards. Increased public awareness and consumer interest led to government involvement in licensure in numerous career areas. By 1980, more than 800 occupations and professions were regulated by law, and performance standards were established by governing boards and agencies (Shimberg, 1982). Assessment of candidates was often part of licensure processes.

During the 1980s, computers began playing a significant role in assessment and measurement. As the capacity and speed of computers increased and costs decreased, they became more available for use in all aspects of the assessment enterprises. Computers played a key role in the development of assessments, the scoring of responses, the analysis of data, and the reporting of results. Item banking, complex analyses and modeling, tailor-made tests, computer-delivered tests and interviews, computer adaptive testing, integrated learning systems, and other developments re-

sulted from improvements in technology, developments in statistical and psychometric theories and methodologies, and ready access to personal computers in schools and workplaces throughout the country.

We also saw some large-scale testing programs that had relied heavily on paper-and-pencil multiple-choice tests start to explore the use of alternative assessment approaches such as essays, observations, and portfolios. In the late 1980s, reform efforts became concerned with national standards. The National Council of Teachers of Mathematics developed standards for evaluating student learning and teacher performance. Other professionals associations have started to develop standards for their fields.

Where Is Assessment Headed?

The 1990s promise several developments that will affect assessment and measurement. One of the more exciting is virtual reality (i.e., computer-generated, visual/aural/tactile, interactive simulations of real-life or imaginary environments and events). Its use, as part of the driver testing process, is being explored in California, Oregon, and Washington. Applicants take a virtual reality form of the driving test before taking the road test. Such a test could also be used to check older drivers for alertness and response time. Surgeons can be tested on the computer, via virtual reality, rather than on cadavers or live patients. The reactions of police officers to emergency situations are being tested by this means. Engineering students use computers to design and modify bridges to optimize structural soundness. These designs are assessed by another computer program. Taxi drivers could be given tests of changes in routes to key locations, based on changes in traffic

conditions, and assessed to see if they optimize travel time while providing good service to the customers. The possibilities are limitless.

Most of the past activities in assessment and measurement had been concerned with performance of individuals. Although we will continue to be concerned with individual performance, we are now also looking at performance of small groups (e.g., cooperative learning groups in classrooms, work groups in the office, teams on construction projects) as well as larger groups and organizations, including governmental agencies. For example, the City of Sunnyvale (CA) has set performance standards for many of its departments, against which they will be assessed. Such standards might include all city streets swept at least twice a month, or 90% of the police calls responded to within six minutes. We anticipate seeing similar practices among other levels of government, including federal agencies under the Clinton-Gore administration's program to "reinvent government."

Why Should You Use This *Resource Handbook*?

Assessment and measurement of performance have been with us for several millennia and continue to increase in importance. It is critical that we develop a working vocabulary, consistent across the many arenas in which assessment and measurement occur. We hope that the "Glossary" section of this *Resource Handbook* will help students, practitioners, and policymakers in various fields move toward a common understanding of such terms. For those who seek more information, the resources listed after the "Glossary" section should be useful.

GLOSSARY

A

Ability: the present or potential competency of an individual to perform a task or to use skills, including ones that are intellectual and physical. [*See* Ability Test, Aptitude, Capacity, Competency, Performance (Individual), Skill, Talent.]

Ability Test: an instrument that estimates the performance level of current or future abilities of an individual. [*See* Ability, Aptitude Test.]

Absolute Standard: a cutting score or designated level of performance that is set without reference to the distribution of performance of individuals in the intended assessment population for which the standard will be used. For example, having an essay test score of 10 on a 12-point scale or doing a pole vault of two-and-a-half meters are absolute standards; scoring in the highest quartile on a writing test or being among the top ten discus throwers in a group of 100 individuals are not absolute standards, but instead relative standards. [*See* Cutting Score, Relative Standard, Standard.]

Accountability: the responsibility for setting appropriate goals; implementing a process, program, policy, or procedure; monitoring and evaluating; producing the specified results or desired outcomes; presenting and interpreting the information and results to others; and justifying decisions made. [*See* Administrative Responsibility, Desired Outcome, Evaluation, Goal, Monitoring.]

Accuracy: the degree to which the data and information collected about the performance of an individual are precise and correct measures of performance, and are free from error. [*See* Bias, Conflict of Interest, Consistency, Credible, Error of Measurement, Precision, Quality Check, Rater Effect, Standard Error of Estimate (SEE), Verification.]

Achievement Test: an instrument designed to estimate what an individual knows about a particular topic or subject area, or to determine the amount of an individual's acquired knowledge and skills in one subject or across subject areas. [*See* Knowledge, Skill.]

Adaptive Testing: an assessment process in which the sequence of items or tasks depends upon responses to previous items or tasks. Performance is assessed on the basis of such characteristics as the difficulty of the items answered correctly rather than on how many items are answered correctly. [*See* Testlet.]

Administrative Responsibility: the accountability, as stated in the agency's policy, for the conduct of the assessment process and for the decisions made

based on the assessment results. [*See* Accountability.]

Adverse Impact: a difference between certain identified groups in the effect of assessment and evaluation results on personnel decisions. The Equal Employment Opportunity Commission's (EEOC) indicator for possible adverse impact is that the selection rate for any race, sex, or ethnic group is less than 80% of the rate for the group with the highest rate. The pool of candidates in which adverse impact is said to have occurred must include only qualified candidates. Although student performance assessment does not result in personnel decisions and thus does not fall under EEOC guidelines, the concept of adverse impact is considered in some cases for assessing the fairness of an assessment to students from certain identified groups. [*See* Fairness.]

Affective Domain: the range of feelings and emotions including interests, attitudes, motivations, values, and appreciations. [*See* Attitude Scale, Cognitive Domain, Psychomotor Domain.]

Alignment: the process of strengthening the linkage among objectives, instructional materials, instructional methods, and assessments, or between job responsibilities and personnel evaluation systems. [*See* Curriculum Alignment.]

Alpha: *See* Coefficient Alpha.

Alternate Forms: two or more comparable, parallel or equated versions of an assessment or measure that

cover the same attributes and are at the same level of difficulty. The measures may be alternate forms and yet differ in format, method of assessment (e.g., paper-and-pencil vs. computer delivery), timing, and number of items. Alternate forms should yield very similar results or scores when taken by the same individual, excluding the effects of additional instruction, coaching, or practice. The three types of alternate forms (comparable, parallel, and equated) differ in the extent of statistical analysis used to compare them. [*See* Comparable Forms, Equated Forms, Parallel Forms.]

Alternative Assessment: a variety of assessment approaches that do not use multiple-choice or closed-response items, but instead require the examinees to generate or produce responses or products. Generally, this includes any assessment technique other than traditional norm-referenced or criterion-referenced paper-and-pencil tests. Examples are essays, portfolios, interviews, observations, work samples, and group projects. [*See* Authentic Assessment, Performance-Based Assessment.]

Amanuensis: one who records the responses or answers for an individual unable to do so due to a physical disability, or who records the responses to an interview.

Analysis: the treatment of data and information in order to elicit certain statistical data, assessment results, and evaluative conclusions. [*See* Congruence Analysis, Data Analysis Plan, Quality Check, Synthesis.]

Analytic Scoring: an approach to scoring or rating that considers various parts or aspects of the attribute or performance being assessed, for use in profiling strengths and weaknesses or in obtaining an overall summary. Scores may be recorded as a check mark for presence or absence of an attribute, marked on a numerical or descriptive rating scale, or put in the form of a brief comment. [*See* Holistic Scoring, Primary Trait Scoring, Scoring, Scoring Rubric.]

Anchor: a referent, such as a common set of items or the same form of a test, to which other measures are linked for comparing or equating, or to provide stability over time in estimating or describing levels of performance. [*See* Benchmark, Common Items, Equated Forms, Equating Drift.]

Anecdotal Record: a short narrative of an event or activity that may be used to document performance or to support generalizations about the performance of an individual or situation.

Appeal Process: a procedure by which the individual who has been assessed or another stakeholder can question the results of an assessment, or can challenge a decision rendered, based upon the results of an assessment. [*See* Stakeholders.]

Appraisal (Performance): *See* Performance Appraisal.

Aptitude: the potential for acquiring abilities or developing competencies. [*See* Ability, Capacity, Competency.]

Aptitude Test: an instrument that estimates an individual's potential for acquiring abilities or developing proficiencies. [*See* Ability, Ability Test.]

Arbitrary: a characteristic of a decision or action that is uninformed or is based upon one person's judgment or discretion. [*See* Capricious.]

Articulated Tests: a series of measures of the same domain designed for use across different ages or grades. The measures include overlapping or adjacent measures of performance in the domain. At each level, objectives and standards should be developmentally appropriate, that is, reflecting the age or grade for which that level is designated. Through vertical equating, performance on difference levels of the tests can be compared. [*See* Battery, Developmentally Appropriate, Equated Forms, Level (Test), Vertical Equating.]

Artifact: (1) that which is artificial, contrived, or fictitious, or which is produced or shaped by an individual. The term can refer to tasks used in and the products of performance assessments (e.g., contrived case descriptions, scale models of objects). (2) an artificial statistical phenomenon or result (e.g., restriction of range, test floor and ceiling, lack of reliability, limited sampling of tasks).

Assess: to estimate the degree of quality or quantity, or to describe or document the nature of an aspect of behavior, learning, or performance. [*See* Estimate *(noun)*, Evaluation, Measure *(verb)*, Rate.]

Assessment: the process or instrument used to measure, quantify, and/or describe those aspects of learning or performance related to the attributes. [*See* Instrument, Measure *(noun)*.]

Assessment Center: the process of using simulation techniques to measure learning, performance, and potential of individuals. This term does not refer to a specific location. Depending on the nature of the assessment, it could be implemented at any of several locations. [*See* Simulation, Surrogate Task, Work Simulation.]

Assessment Plan: the overall structure for the implementation of an assessment program including, but not limited to, areas to be assessed, individuals to administer the assessments, locations and schedules for administration, production or acquisition and storage of materials and equipment, identification and registration of test takers, and scoring and reporting timelines. [*See* Formal, Informal, Schedule, Timeline.]

Assessor: the person who collects data and who measures attributes related to the performance of an individual. Assessors may be principals, teachers, supervisors, other students, parents, management staff, other employees, government inspectors, board members, or community representatives. This term includes, but is not limited to, the interviewer, judge, observer, and scorer. [*See* Evaluator, Interviewer, Judge, Observer, Scorer.]

Attainable: that which is feasible and possible for an individual or group to reach or be able to do. [*See*

Developmentally Appropriate, Feasibility (of Performance Assessment), Teachable.]

Attitude Scale: a measure on which individuals rate or provide data on their feelings, beliefs, values, and opinions. [*See* Affective Domain, Attribute, Likert Scale, Rating.]

Attribute: a characteristic, behavior, skill, knowledge, or perceived quality such as, but not limited to, attitude, ability, or interest. [*See* Construct.]

Audience: those individuals, clients, or consumers who have a potential interest in the results of student and/or employee performance assessment and in the quality of learning and of performance. [*See* Stakeholders.]

Audit: an independent quality check and verification of the assessment process and results. [*See* Monitoring, Quality Check, Verification.]

Authentic Assessment: an assessment approach that has been designed to provide a realistic task, activity, simulation, or problem related to that attribute or performance being measured. [*See* Alternative Assessment, Performance-Based Assessment, Portfolio, Simulation, Surrogate Task, Work Simulation.]

Average: an estimate of typical or central performance across individuals in a group or, for an individual, over a period of time. Numerical averages usually refer to the mean. [*See* Central Tendency Measure, Mean, Median, Mode.]

B

Basic Skills Test: a measure of fundamental skills in reading, mathematics, and/or writing. Such tests are often used to determine if a student has acquired a minimally acceptable level of attainment of skills in these three areas needed to fulfill exit requirements from a level of schooling or to move to the next level of schooling. Basic skills tests are sometimes used for hiring decisions when such skills are needed for minimally acceptable performance on the job.

Battery: a set of tests on different areas or attributes that have been standardized on the same sample so that results on different tests in the battery can be used together for decision making. [*See* Articulated Tests, Standardization Sample.]

Behavior-Based Assessment: a process for measuring performance that focuses on specific and observable actions of an individual. [*See* Behavioral Checklist, Behaviors, Knowledge-Based Assessment, Outcomes-Based Assessment, Performance-Based Assessment, Skills-Based Assessment, Spontaneous Performance Assessment, Structured Performance Assessment.]

Behavioral Checklist: an instrument on which the assessor indicates the presence or absence of specific behaviors (e.g., child can button shirt, student can count to 20 in French, employee can replace toner

in the copier). The checklist may be designed to assess both spontaneous and structured performance. [*See* Behaviors, Checklist, Spontaneous Performance Assessment, Structured Performance Assessment.]

Behaviorally Anchored Rating Scale (BARS): a type of assessment in which judgments about performance are empirically linked to specific examples of incumbent performance at each level of effectiveness on the rating scale. [*See* Rating, Scale (Rating).]

Behaviors: the actions of individuals or groups that are specific and observable.

Benchmark: a referenced behavior for comparing an observed performance at a given level. [*See* Anchor, Holistic Scoring, Objectives, Range Finders, Rating, Scoring Rubric.]

Bias: **(1)** a systematic tendency toward a lack of objectivity, fairness, or impartiality on the part of the assessor or evaluator, often based on personal preferences and inclinations. **(2)** systematic error in the assessment instrument or procedures, or in the interpretation and evaluation process. [*See* Contamination, Culture Fair Testing, Culture Free Testing, Differential Functioning, Error of Measurement, Fairness, Rater Effect.]

Biserial Correlation: *See* Correlation, Point-Biserial Correlation, r-Biserial Correlation.

Blueprint: *See* Content/Process Matrix, Specifications (Assessment).

Branching Test: *See* Adaptive Testing.

Breadth and Depth: the scope and extent of detail of the domain or content to be assessed. Alternative assessments are apt to sacrifice breadth of coverage for depth in the area assessed, whereas multiple-choice tests tend to emphasize breadth rather than depth. [*See* Sampling of Content, Scope of Content.]

C

Capacity: the potential for acquiring skills and competencies through such means as instruction, coaching, practice, self-study, or training. [*See* Ability, Aptitude, Competency, Skill.]

Capricious: that which is unpredictable, unaccountable, impulsive, or without a supporting rationale, or which is characterized by implied unruly or whimsical behavior. [*See* Arbitrary.]

Ceiling: the highest limit of performance that can be assessed or measured by an instrument or process. Individuals who can perform near to or above this upper limit are said to have reached the ceiling; these individuals should be given a higher or more difficult level of the instrument or process in order to obtain a valid estimate of their performance. If the ceiling of an instrument is too low for a group of individuals, the distribution of their scores will tend to be negatively skewed. [*See* Floor, Functional-Level Testing, Level (Test), Skewness.]

Central Tendency Effect: a type of rater effect in which an assessor or an evaluator tends to rate individuals or groups toward the mid-point of a scale or to judge the performance as average or neutral when it is actually well above or well below the middle level of the scale. This use of the term "central tendency" is not the same as that used in

statistics. [*See* Central Tendency Measure, Rater Effect.]

Central Tendency Measure: one of several statistical summary values of typical performance, including the mean, median, and mode. [*See* Average, Mean, Median, Mode.]

Essay Scores
6
6
5
5
5
4
4
4
4
3
3
2
1
1
1
$\Sigma = 54$
$N = 15$

Mean = arithmetic average = $\Sigma \div N = 54 \div 15 = 3.6$

Median = middle score, which in this case is the eighth score from the top or bottom = 4

Mode = score with the highest frequency = 4

Chance-Level Score: assuming that test takers complete all items on a test, the expected level of performance or score that would be obtained through purely random guessing. An estimate of the chance-level score for a multiple-choice test is the total number of items on the test divided by the

number of answer choices or options per item. For example, a chance-level score on an 80-item test with five choices per item is 80 divided by 5, or 16. For a 35-item test with four choices per item, a chance-level score would be 9. [*See* Chance Score.]

Chance Score: for an individual, the chance score is an estimate based on how many items that individual attempted to answer, normally considered the number of items for which an answer has been given, or the last item answered, counting both skipped items and attempted items. For example, if on an 80-item, four-choices-per-item test, the last item marked on the sheet was item 64 and the individual omitted no items, the chance score for that individual would be 64 divided by 4, or 16. If skipped items are considered "not attempted items," this individual's chance score would be 64 minus 10 non-attempts, or 54 divided by 4, or 14 (i.e., 13.5 rounded off). A score of 20 may reflect more than random guessing if an individual does not respond to several items or reach the end of the test. [*See* Chance-Level Score.]

Cheating: to act in a dishonest manner in order to obtain a higher level of performance on an assessment. Examples include having prior access to the test materials, copying from another test taker, or the use of an impostor. [*See* Coaching, Test (Score) Pollution, Test Security, Test Wiseness.]

Checklist: an instrument that specifies criteria or indicators of merit and on which the assessor or evaluator marks the presence or absence of the attribute being assessed. [*See* Behavioral Checklist.]

Classical Test Theory: a model of test score analysis that focuses on characteristics of the items and test, but not on characteristics of the examinees from which the item and test data were obtained, and that assumes that an individual's obtained score is the sum of the true score and the error score. [*See* Error of Measurement, Generalizability Theory, Item Response Theory (IRT), Obtained Score, True Score.]

Classification Error: the incorrect categorization of individuals or groups, based on using obtained scores as compared to their hypothetical true scores and due to the fallibility of assessments as predictors. [*See* Error of Measurement, False Negative, False Positive, Obtained Score, True Score.]

Classification Rate: the percentage of individuals assigned to various categories or classifications, based on assessment results (e.g., 82% of students passed English class, 87% of employees knew how to operate a new piece of equipment safely).

Clinical Interviews: those interviews conducted by psychologists, social workers, counselors, therapists, and similar professionals to obtain information relevant to the performance of an individual or a group in a learning or work setting. [*See* Interview.]

Coaching: the assistance given to individuals in ways that will improve their performance, not just for the purpose of doing better on the assessment, but also for the purpose of improved performance and increased learning. Coaching may include reviewing materials in the areas covered by the assessment,

providing on-the-job training or tutoring for the attributes being assessed, giving encouragement and building confidence, offering feedback on strengths and weaknesses, giving practice tests, and teaching about test-taking strategies. [*See* Mentoring, Practice Test, Test (Score) Pollution, Test Wiseness.]

Code: a symbol, either numeric or alphabetic or a combination, used to represent attributes, events, or words (e.g., 3 = Third Grade, M = Math, TQ = Teacher asks question, AnTl = Answers telephone, 14 TE = 14 typing errors).

Coding: the process of transforming data, evidence, information, judgments, notes, and responses to numeric and/or alphabetic codes. [*See* Code.]

Coefficient Alpha: an internal-consistency estimate of reliability, used not only when items are scored right/wrong, but also with essay scores, task ratings, and attitude scales. Sometimes this term is referred to as Cronbach's Alpha or simply Alpha. [*See* Kuder-Richardson 20 (KR20) Formula, Reliability.]

Cognitive Complexity: the individual's use of multiple skills and processes when solving assessment tasks as opposed to simpler, less developmentally advanced ways to solve the problems in a specific content domain. No assumptions can be made about the cognitive complexity of an assessment without analyzing the specific skills and processes that individuals actually employ when doing the assessment task. [*See* Cognitive Domain, Higher-Order Thinking Skills (HOTS), Taxonomy (Cognitive Domain).]

Cognitive Domain: the range of knowledge and knowledge-related skills needed for individuals to achieve different types of learning objectives. These range from perception to knowledge of facts and acquisition of skills to higher-order inference. [*See* Affective Domain, Higher-Order Thinking Skills (HOTS), Psychomotor Domain, Taxonomy (Cognitive Domain).]

Combined Model: a scoring or evaluation procedure that uses features of both compensatory and conjunctive models. [*See* Compensatory Model, Conjunctive Model.]

Comments: the information provided by anyone involved in the assessment process (e.g., individual being assessed, teacher, supervisor, observer, interviewer, data collector, proctor, judge, analyzer, evaluator) concerning incidents or factors that could affect the quality and accuracy of the assessment data and the judgments and evaluations made about performance levels (e.g., fire alarm sounded during the writing sample task, power failed during the word processing test, supervisor was called out for 15 minutes during the interview on her performance to take an important telephone call concerning next month's layoffs in her department). [*See* Critical Information, Notes.]

Common Items: a set of identical test questions that appears in two or more forms of a test for use in equating the forms. These questions may be scattered throughout the two forms or kept together as a block in approximately the same locations in each form of the test. [*See* Anchor, Equated Forms, Form (Test).]

Comparability: the similarity of phenomena (e.g., attributes, performances, assessments, data sources) being examined, but not based on a statistical analysis or the statistical characteristics of the two phenomena. The amount or degree of comparability is often used to determine the appropriateness of using one phenomenon in lieu of another and to help ensure fairness. [*See* Comparable Forms, Equivalence, Fairness.]

Comparable Forms: two or more assessments that appear to be similar to one another, based on an examination of their purpose and contents, but for which there are no statistical data indicating the degree of similarity. [*See* Alternate Forms, Comparability, Equated Forms, Parallel Forms.]

Compensatory Model: an evaluation or scoring procedure that permits the trade-off of one attribute against another (i.e., low performance on one attribute can be offset by high performance on another). Most compensatory models incorporate features of a conjunctive model by having an absolute minimal level of performance for each attribute, below which trade-offs are not permitted; such cases are referred as "combined models." [*See* Combined Model, Conjunctive Model, Decision Rules, Disjunctive Model.]

Competence, Competences: a repertoire of competencies. [*See* Competency, Merit.]

Competency, Competencies: a knowledge, skill, ability, personal quality, experience, or other characteristic that is applicable to learning and success in

school or in work. [*See* Ability, Competence, Competency-Based Assessment, Knowledge, Skill.]

Competency-Based Assessment: a process for measuring performance that focuses on an individual's competencies. [*See* Competency.]

Competency Test: an instrument that assesses whether an individual has attained a minimum standard of skills and knowledge, and should be promoted, graduated, licensed, or hired. [*See* Competency, Proficiency Test.]

Complex Reporting Scheme: *See* Reporting Scheme (Complex).

Component: one of the parts or processes in an assessment system (e.g., group project, writing sample, portfolio, multiple-choice test).

Composite Score: a score that combines two or more scores or results for the same or related attributes.

Computerized Assessment: the use of computers to measure performance on some attribute, not necessarily an attribute related to computers and technology.

Concurrent Validity: *See* Validity.

Conference: a meeting between the teacher and the student and/or parents, or the supervisor and employee, to discuss mutual concerns and to promote an understanding of the assessment process and re-

sults, including the instruments being used and the criteria and standards being applied. [*See* Debriefing Interview, Interview.]

Confidence Interval: the upper and lower limits of an obtained set of values or scores, within which the true value or score, or a critical score such as a passing score, is likely to occur a certain percentage of the time. The designated confidence level determines how many standard errors of measurement (SEMs) to use in calculating the confidence limits. Percentile bands are one commonly used types of confidence interval in testing. [*See* Error of Measurement, Obtained Score, Percentile Band, Precision, Standard Error of Measurement (SEM), True Score.]

Confidentiality: the protection of data and information from persons other than those authorized to have access to them. [*See* Consent, Informed Consent, Reporting.]

Configural Scoring Rule: a rule for interpreting a pattern of scores on two or more assessments or parts of one assessment for the same individual. [*See* Synthesis, Weighted Score.]

Conflict of Interest: a situation in which the private interests of someone involved in the assessment or evaluation process (e.g., teacher, supervisor, scorer) are incompatible with that person's role in the process, thereby potentially compromising the quality of the assessment activities, the accuracy of the data, and/or the results of the assessment or evaluation. [*See* Accuracy.]

Example of Calculating Confidence Intervals

Scaled Score	Percentile	Scaled Score (rounded) Confidence Interval for 116 68%	95%	99%	Percentile Band for 116 68%	95%	99%
...	...						
123	55						
122	54						
121	53						
120	52						
119	51						
118	50						
117	50						
116	48	114-118	112-120	109-123	45-50	40-52	30-55
115	47						
114	45						
113	43						
112	40						
111	37						
110	33						
109	30						
...	...						

SEM = 2.2

For 68% confidence interval, ± 1 SEM; for 95% confidence interval, ± 2 SEMs; for 99% confidence interval, ±3 SEMs.

± 1 SEM	± 2 SEMS	± 3 SEMS
116 - 2.2 = 113.8 ~ 114	116 - 2(2.2) = 111.6 ~112	116 - 3(2.2) = 109.4 ~109
116 + 2.2 = 118.2 ~118	116 + 2(2.2) = 120.4 ~120	116 + 3(2.2) = 122.6 ~123

Congruence Analysis: the verification of data by using more than one instrument or source of data for assessing performance on the same criterion. [*See* Multiple Measures, Triangulation, Verification.]

Conjunctive Model: an evaluation or scoring procedure that requires an individual to attain a minimal level of performance on all attributes assessed. [*See* Combined Model, Compensatory Model, Decision Rules, Disjunctive Model.]

Consent: the granting of permission by an individual, or the person responsible for that individual (e.g., parent), concerning the collection, use, retention, or access to assessment data and information. [*See* Confidentiality, Informed Consent, Reporting.]

Consequences: any outcomes that occur as a result of implementing an assessment or measurement process. For example, students check out library books related to the topics to be covered on the final examination; the addition of a time deadline for feeding convalescent patients to the assessment checklist results in the nurses aides reallocating their time to meet the new mealtime deadlines; performance of the twelve automobile mechanics indicates that only five have minimally acceptable levels of performance in brake repair and a training program for the other seven mechanics is implemented as a consequence of the assessment. [*See* Unintended Consequences, Validity--Consequential Basis of.]

Consequential Basis of Validity: *See* Validity.

Considered Necessary: that which is judged to be required, but may not be sufficient, to obtain an accurate and valid estimate of performance or to make a sound decision.

Consistency: **(1)** an attribute of a test or assessment that indicates whether an individual's score will be nearly identical if measured on two or more occasions with no interventions in between testing. **(2)** the degree to which a judge or different judges scoring the same test or assessment agree across test takers and/or over time. **(3)** the degree to which all items on a test measure the same attribute. [*See* Accuracy, Inconsistency, Precision, Rater Drift, Rater Effect, Reliability.]

Construct: an attribute of an individual or a phenomenon that is not directly observable, but which is theoretically-based or is inferred from empirical evidence (e.g., sales staff enthusiasm about a new product being marketed is inferred from the number of new clients contacted; a student's poor attitude toward reading is inferred from the number of missing or incomplete homework assignments that require reading; the cohesiveness of an office is inferred from the numbers and types of helping behaviors among staff). [*See* Attribute, Operational Definition, Operationalize, Validity--Construct.]

Construct Validity: *See* Validity.

Constructed Response Test: a type of performance assessment comprised of free-response or open-ended items. Individuals generate answers to the assessment items rather than selecting answers from sets of possible responses, as is the case with

multiple-choice tests. [*See* Multiple-Choice Test, Open-Ended Extended Item, Short-Answer Item.]

Contamination: a tendency for the assessor's data, the scorer's ratings and judgments, or the evaluator's conclusions to be influenced or confounded by irrelevant knowledge about the individual or other factors that have no bearing on the individual's level of performance. [*See* Bias, Error of Measurement, Rater Effect.]

Content Validity: *See* Validity.

Content/Process Matrix: a matrix, or blueprint, used in planning the development of a test. After a list of objectives are specified and the content to be covered by a test has been developed, a matrix is constructed with one dimension being the specific content to be covered and the second dimension being the processes used to solve the problems. For example, in a biology class, students may be expected to demonstrate that they know the difference between frogs and toads (content) by drawing and labeling an example of each type of amphibian (process). Or a chemist might be asked to compare (process) four samples of drinking water for levels of lead (content). The content/process matrix is used to classify each assessment item. Each cell in the matrix can be assigned a weight, in terms of its importance, for use in calculating an individual's final score. [*See* Skills Hierarchy, Specifications (Assessment).]

Context (Learning): the environment within which the individual learns. This includes, but is not limited to, the physical facilities and settings of the

Example of a Content / Process Matrix

	Content		
Process	Fractions	Decimals	Percents
Calculating			
Converting			
Estimating			
Comparing			
Problem Solving			

classroom or workplace, the school or business, the community, and the individual's home; the types of individuals in the peer group; resource availability (staff, materials, equipment, time); and the demographic characteristics of classrooms, schools, work settings, and communities, including their geographic location, and their ethnic, linguistic, racial, and gender compositions.

Contextual Variables: *See* Notes.

Contract: a binding agreement (written or oral) between two or more parties concerning expectations, responsibilities, and possibly remuneration or award of each party.

Contrast Effect: a type of rater effect in which an assessor or an evaluator tends to compare one individual or group to other individuals or groups rather than comparing that individual's or group's level of performance to the standard(s). [*See* Rater Effect.]

Conversion Parameters: the numerical formulae or quantitative procedures for transforming results on

one assessment to results on an equivalent form of that assessment. [*See* Equating, Parameter.]

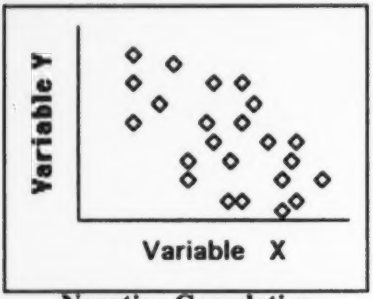
Correction for Guessing: the application of a formula to adjust the raw score earned on a multiple-choice test for random guessing. The formula score usually represents the number of items answered correctly minus a percentage of the items answered incorrectly (e.g., number right minus one-quarter the number wrong). [*See* Derived Score, Multiple-Choice Test, Raw Score.].

Correlation: the degree of relationship (linear or curvelinear) between two variables, scores, or assessments. Correlations range from -1.00 to +1.00. A negative value indicates that as the value of onevariable increases the other tends to decrease. A positive value indicates that as one increases in value so does the other and that as one decreases so does the other. A correlation of zero indicates no statistical relationship between the two variables. [*See* Point-Biserial Correlation, r-Biserial Correlation, Scatterplot, Validity Coefficient.]

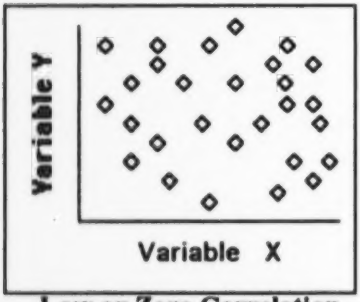
Corroborating Evidence: the documentation that confirms or strengthens, and that provides support for other documentation on the same attribute, competency, or situation. [*See* Evidence, Triangulation.]

Cost: the quantity of resources required in order to achieve a desired end. The cost of a performance assessment involves not only the cost of materials, equipment, staff, and space for the administration,

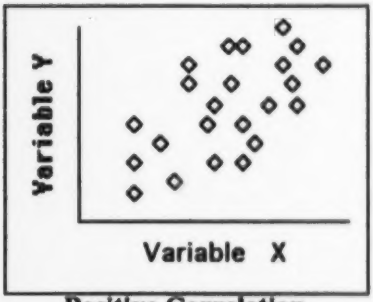
Three Types of Relationships for Correlations



Negative Correlation



Low or Zero Correlation



Positive Correlation

but also such factors as time, disruption, foregone opportunities, energy, and stress that are not completely reducible to monetary terms.

Credible: that which is worthy of confidence and acceptance by others, usually based on the expertise, trustworthiness, and/or reliability of the source of the evidence or judgment. Credible does not necessarily mean accurate or valid. [*See* Accuracy, Reliability, Validity--Face.]

Criterion, Criteria: (1) an external basis upon which performance (e.g., learning, acquisition of a skill) is rated or judged as successful or meritorious. Each criterion falls within a **domain** covered by the assessment or evaluation system and is defined by **elements**, **indicators**, and **descriptors** (see below for examples). The indicators and descriptors should be stated specifically and in measurable or observable terms. Satisfactory levels of performance on criteria are specified by standards. (2) a variable that is predicted from one or more other variables. [*See* Foundation, Standard, Validity--Criterion Related, Variable.]

Domain: a broad area having some common characteristics and for which criteria and standards are specified for assessing performances in that domain.

Element: a category of knowledge, skills, abilities, behaviors, and attributes that falls within a domain.

Indicator: for each element, the types of knowledge, skills, abilities, behaviors, and attributes that are empirically or by definition connected to the criterion.

Descriptor: for each indicator, a specific example of the performance being assessed.

Criterion-Referenced Test (CRT): an instrument developed and used to estimate how much of the content and skills covered have been acquired by the individuals assessed. Performance is judged against a set of criteria rather than in comparison to other individuals tested, as with norm-referenced tests, or to expertise in a broad knowledge area, as with domain-referenced tests. [*See* Domain-Referenced Test, Norm-Referenced Test (NRT).]

Criterion-Related Validity: *See* Validity.

Critical Incident: a significant and observable episode or performance (effective or ineffective) that alters the direction of subsequent events, behaviors, or activities. [*See* Critical Incident Appraisal.]

Critical Incident Appraisal: the use of documentation concerning critical incidents when evaluating and making decision about an individual's current and potential performance. Such documentation may include information on noteworthy accomplishments, substantive improvement, and/or significant failures. [*See* Critical Incident, Documentation.]

EXAMPLES FOR CRITERIA

Example One: Criterion - High school students can analyze historical events in major regions of the world.

Domain - I. Knowledge of Middle Eastern Countries

Elements -
I.A. Natural Resources of Middle Eastern Countries
I.B. Political History of Middle Eastern Countries

Indicators -
I.A.1. Students know the major imports and exports of Middle Eastern countries.
I.A.2. Students are aware of the impact of processing resources on the environment in Middle Eastern countries.
I.A.3. Students know how the availability of certain natural resources affects art, architecture, and other aspects of culture in Middle Eastern countries.
I.B.1. Students know the history of Middle Eastern economic development.
I.B.2. Students know the major political leaders in today's Middle Eastern countries.
I.B.3. Students know the religious movements in the Middle East and their impact on the political history of Middle Eastern countries.

Descriptors -
I.A.1.a. Students mark correctly three major exports of each of five different Middle Eastern countries on a map.
I.A.2.a. Students prepare a notebook of articles showing five problems for environmental impact of oil refineries and oil transport systems in the Middle East.
I.A.3.a. Students write a report including several examples of how Islamic decorative art utilized the available natural resources in the years 500 to 1300 A.D.

I.B.1.a. During an oral analysis of a news story on the Middle East, students are able to distinguish between the most and the least industrialized countries in the Middle East.

I.B.2.a. Students write brief biographies on three current Middle Eastern leaders, pointing out their stand on three major political issues.

I.B.3.a. Students pass a multiple-choice test (three-fourth of answers correct) about the rise of Islam in the 7th century A.D. and its spread throughout parts of Asia, north Africa, and southern Europe.

Example Two: Criterion - The carpenter has the skills to use tools of the trade in an appropriate and safe manner.

Domain - I. Use of Tools

Elements -

I.A. Skill in Using Different Types of Saws

I.B. Use of Safety Equipment and Practices

Indicators -

I.A.1. Individual uses several types of hand saws (e.g., coping, crosscut) and power saws.

I.B.1. Individual uses goggles and ear plugs when using power saws and places hands in a safe position.

Descriptors -

I.A.1.a. Individual cuts a board at a 45-degree angle, using a hand crosscut saw.

I.A.1.b Individual uses circular power saw to cut and trim boards for hardwood floors to the correct lengths.

I.A.1.c. Individual uses a power coping saw to cut the curved corner of a counter top.

I.B.1.a. Individual wears safety glasses with side extenders when using circular power saw.

I.B.1.b. Individual's placement of hands is done in a safe manner when using crosscut saw.

I.B.1.c. Individual wears the ear protectors when operating power coping saw.

Critical Information: the knowledge about the individual being assessed or the assessment instruments and conditions that must be known to avoid incorrect interpretations of results. [*See* Comments.]

Critical Score: *See* Cutting Score, Passing Score.

Cronbach's Alpha: *See* Coefficient Alpha.

Culture Fair Testing: the use of assessment processes and materials designed to minimize bias in favor of or against certain groups. [*See* Bias, Culture Free Testing, Fairness.]

Culture Free Testing: the use of assessment processes and materials that claim to be uninfluenced by cultural factors and therefore equally fair to all groups that might take the test or be affected by the assessment. The possibility of culture free testing is questionable. [*See* Bias, Culture Fair Testing, Fairness.]

Curricular Validity: *See* Validity.

Curriculum: **(1)** a comprehensive overview, including activities planned for delivery to the students, the scope of content, the sequence of materials, interpretation and balance of subject matter, and motivational, instructional, and assessment techniques to be used. **(2)** a set of ordered, intended learning outcomes.

Curriculum Alignment: altering the curriculum to match the breadth (scope), depth (detail), and se-

quence of a testing or assessment program. This practice ensures that students have been or will have been exposed to those content areas being assessed prior to testing. [*See* Alignment, Curriculum, Narrowing of the Curriculum, Teaching to the Test.]

Curve: *See* Grading on the Curve, Normal Curve.

Cutting Score: a score that marks the difference between two levels of performance (e.g., good and excellent). When the difference is between minimally acceptable and not acceptable, or pass and not pass, it is referred to as a cut score, critical score, or passing score. [*See* False Negative, False Positive, Passing Score, Standard, Standard Setting.]

D

Data: the information and evidence gathered during the assessment process for use in determining the level of performance. [*See* Extant Data.]

Data Analysis Plan: a delineation of the steps to be undertaken in the processing, quality control, and analysis of data. The plan should indicate which data are to be included, how data should be checked for quality control purposes, which formulae or statistical programs should be applied to the data, in what format the data should be summarized (e.g., distribution tables, bar graphs, scatterplots), and for what decisions the data are to be used. In addition, persons responsible for each step, a timeline, and a list of resources needed should be included in the plan. [*See* Analysis, Data Collection Procedures, Procedures (Assessment), Quality Check.]

Data Collection Procedures: the steps and sources used to obtain qualitative and quantitative data and information about an individual's background characteristics and performance. [*See* Data Analysis Plan, Procedures (Assessment).]

Data Integration: the merging of related data for use in scoring, judging, evaluating, and decision making. [*See* Synthesis.]

Data Sources: the persons, documents, products, activities, events, and records from which the data are obtained.

Debriefing Interview: a series of questions asked after an assessment activity to clarify behaviors and activities, to fill in missing information, to encourage an individual to reflect on his/her performance, and to identify future directions and plans for improving performance. [*See* Conference, Exit Interview, Interview.]

Decile: one of nine points that divide a distribution into ten equal parts. For example, the 2nd decile has one fifth of the distribution below it and four fifths above it. The 2nd decile corresponds to the 20th percentile, the 5th decile to the 50th percentile, and so on. [*See* Normal Curve, Percentile.]

Decision Rules: the guidelines for determining the level of merit, value, or worth of an aspect of an individual's performance from available, relevant data and information, or for connecting those performance levels with subsequent actions such as passing a course, being hired, or being promoted. [*See* Compensatory Model, Conjunctive Model, Disjunctive Model, Merit, Value, Worth.]

Defensible: a conclusion or statement that is explainable or justifiable, based on a solid foundation and policy, explicit procedures, valid assessments, and sound evaluation and decision-making practices.

Demonstration (Performance): a specified assessment task, procedure, or activity that the individual

being assessed implements and during which performance of that individual is observed. For example, a first aid trainee shows how to splint a broken leg on another trainee while being observed by the instructor. [*See* Exhibition, Exhibition of Mastery Principle.]

Derived Score: a score to which the raw score has been transformed or converted. [*See* Formula Score, Grade-Equivalent Score (GE), Normal Curve Equivalent (NCE), Percentile Rank (PR), Raw Score, Scaled Score, Stanine, T-score, z-score.]

Descriptor: *See* Criterion, Scoring Rubric.

Desired Outcome: the results or products that a training program, process, instructional unit, or learning activity strives to achieve, as defined in measurable terms. [*See* Goal, Objectives.]

Developmentally Appropriate: a characteristic of those items or assessment tasks that reflect what individuals at a particular age and cognitive level are trying to learn or what skills they are trying to acquire and which they have a reasonable chance of being able to learn or acquire. Developmentally appropriate items and tasks used in performance assessments should consider the complex nature of how people learn, including the understanding that knowledge and skills evolve with increased experience and physical and mental development. [*See* Articulated Tests, Attainable, Functional-Level Testing, Teachable.]

Diagnosis: the process of determining an individual's strengths and weaknesses, based on the results of the assessments. Diagnosis is an essential preliminary process to preparing an instructional plan for a student or a development plan for an employee in those cases where remediation or special services are needed. [*See* Remediation.]

Dichotomous: a characteristic of a variable that can be grouped or classified into two mutually exclusive categories (e.g., female/male, pass/fail, present/absent, right/wrong). [*See* Variable.]

Difference Score: the amount of change (increase or decrease) for an individual or a group (based on its mean score) between the most recent score and a prior score on the same or an equivalent test (also called a "gain score"), or the amount by which the results on two assessments differ from each other. Difference scores should be calculated using scaled scores, stanines, or normal curve equivalents. Since raw scores and percentile ranks are not equal-interval scales, they should not be used for such calculations. [*See* Normal Curve Equivalent (NCE), Percentile Rank (PR), Raw Score, Regression Effect, Scaled Score, Stanine.]

Differential Functioning: a characteristic of an assessment approach, test item or task, instrument, or evaluation system that yields higher results for one group than another group, even though both groups have the same level of ability or competence on that aspect of performance. [*See* Bias.]

Differential Item Functioning (DIF): *See* Differential Functioning.

Differential Prediction: the extent to which a measure estimates future performance on the same attribute differently for two or more groups of people who vary on relevant characteristics (e.g., number of years of mathematics courses for a mathematics achievement test).

Discrepancy: the difference between two or more raters or scorers on the same assessment, or between two or more evaluators concerning the same evaluee. [*See* Evaluee, Rater Effect, Reliability.]

Discrimination (Item): the extent to which an item, task, or test differentiates between individuals with different characteristics. [*See* Item Analysis.]

Disjunctive Model: an evaluation or scoring procedure that requires the individual to achieve a minimal level of performance on only one of the attributes being assessed. [*See* Compensatory Model, Conjunctive Model, Decision Rules.]

Distractors: the alternatives or options in a multiple-choice item that are not the correct or best choice for a response. [*See* Multiple-Choice Test, Stem.]

Documentation: the collection or compilation of all tangible materials, records, and forms used in the assessment and evaluation of an individual's performance. [*See* Record *(noun)*.]

Documenting: the process of recording and providing tangible evidence and information about the performance of an individual or group.

Domain: *See* Criterion.

Domain-Referenced Test: an instrument developed and used to estimate the degree of proficiency or level of expertise possessed by the individuals being assessed. Performance is judged against desired levels of expertise or proficiency in a broad area or domain rather than against a set of criteria, as with criterion-referenced tests, or in comparison to other individuals tested, as with norm-referenced tests. [*See* Criterion-Referenced Test (CRT), Domain Score, Norm-Referenced Test (NRT).]

Domain Score: an estimate of the proportion of a knowledge, skill or ability domain that an individual has mastered (e.g., percent of 2-digit addition problems answered correctly). Domain scores are most meaningful when a domain is rigorously defined. [*See* Domain-Referenced Test.]

Drift: *See* Equating Drift, Rater Drift.

E

Effectiveness: an attribute of those individuals, agencies, programs, services, and approaches that meet their stakeholders' needs. [*See* Efficiency, Stakeholders.]

Efficiency: an attribute of those individuals, agencies, programs, services, and approaches that balance effectiveness against considerations of costs (i.e., are effective with a minimum use of resources). [*See* Effectiveness.]

Element: *See* Criterion.

Enhanced Multiple-Choice Item: a type of multiple-choice item that involves problem-solving and that may require several minutes to determine which choice is correct. [*See* Multiple-Choice Test, Open-Ended Extended Item.]

Environment: *See* Context (Learning).

Equated Forms: comparable or parallel forms, measures, or scores that have been statistically linked to each other and can be used in place of each other. [*See* Alternate Forms, Comparable Forms, Equating, Equipercentile Equating, Equivalence, Horizontal Equating, Parallel Forms, Vertical Equating.]

Equating: the statistical process of linking two or more forms of the same assessment so that the results on each form can be transferred to a common score scale and used in place of each other. [*See* Conversion Parameters, Equated Forms, Equating Drift, Equipercentile Equating, Horizontal Equating, Vertical Equating.]

Equating Drift: the systematic change in successive forms of a test over a period of time in the levels of scale-referenced scores that is due to the equating process, including which anchors are used. [*See* Anchor, Equated Forms, Equating.]

Equipercentile Equating: a technique that regards two performance estimates or scores as equivalent if they have the same percentile rank in two distributions based on the same or comparable samples. [*See* Equated Forms, Percentile Rank (PR).]

Equitable: that which is fair, impartial, and just, and which provides equal opportunity for all. [*See* Fairness.]

Equity: a quality or state that is fair, impartial, and just. [*See* Equitable.]

Equivalence: the comparability of two or more parallel measures that have been designed to assess the same aspect of performance and to yield similar results regardless of the measure used or the scoring/rating procedures applied (e.g., two different problems on the same geometric shape; two different songs to sight read; two different blood tests to

be performed in the hospital lab). [*See* Comparability.]

Error of Measurement: the difference between an individual's obtained score and his/her true score on an assessment that is due to factors inherent in the assessment, including lack of reliability in the assessment instrument or process, variability in the administrative procedures or settings for the assessment, limited sampling of the individual's performance within the domain assessed, rater effects, and interactions among such factors. Error of measurement is also called error score. [*See* Assessor, Bias, Classical Test Theory, Confidence Interval, Contamination, Estimate *(noun)*, False Negative, False Positive, Obtained Score, Precision, Rater Effect, Reliability, Residual Score, Sampling Error, Sampling of Content, Sampling of Performance, True Score.]

Error Score: *See* Error of Measurement.

Essay: a narrative (written or oral) description, analysis, explanation, interpretation, opinion, and/or summary that demonstrates an individual's use of information and language in generating a coherent document about a particular topic or in response to a question. [*See* Prompt.]

Estimate *(noun)*: an approximation of a true score, parameter, or value. For example, an obtained score on a physics achievement test is an estimate of the true score and of the test taker's knowledge of physics. A Coefficient Alpha is an estimate of a test's internal consistency. A rating of a portfolio is an estimate of the quality of that portfolio and the

performance of the individual it represents. Because no instrument or statistical procedure can provide an exact (or true) score or value for an individual, an instrument, or a phenomenon, essentially all data are estimates. The smaller the error of measurement, the more precise the estimate of the true score or value. [*See* Data, Error of Measurement, Obtained Score, Parameter, Precision, Residual Score, Score, Standard Error of Estimate (SEE), Standard Error of Measurement (SEM), True Score.]

Evaluation: the systematic process of determining or judging the merit, value, or worth of someone (the evaluee, such as a teacher, student, or employee) or something (the evaluand, such as a product, program, policy, procedure, or process). [*See* Merit, Performance Appraisal, Value, Worth.]

Evaluator: a person who assembles data and information about an individual (the evaluee), analyzes them, makes judgments as to whether that individual's performance level meets the pre-specified standards, prepares a summary report, writes recommendations, and may provide feedback to the individual or others, directly or through another person. In general, evaluators determine the overall merit, value, and/or worth of the evaluee. [*See* Assessor, Evaluee, Judge.]

Evaluee, Evaluatee: the person whose qualifications and performance are being evaluated. [*See* Evaluator.]

Evidence: the documentation and verbal statements by credible witnesses (e.g., supervisors, peers, clients)

concerning the knowledge, skills, abilities, and behaviors of an individual. Evidence is used both to generate and to justify judgments about a individual's performance for purposes of evaluation. [*See* Corroborating Evidence, Credible.]

Evidential Basis of Validity: *See* Validity.

Excellence: a quality or state of high or superior performance, or of having virtues and value surpassing others.

Exemplary Performance: a level of performance that is regarded as worthy of imitation or modeling. [*See* Meritorious Performance, Modeling.]

Exhibition: an inclusive demonstration of skills or competencies. The performance is judged against standards of excellence known to participants ahead of time. Typically these are live performances, often interdisciplinary, and may require individual creativity as well as a display of developed skills. Exhibitions include competitions among individuals (e.g., debates) or individual or group/collaborative projects produced over an extended period of time. [*See* Demonstration (Performance), Rites of Passage.]

Exhibition of Mastery Principle: a demonstration of competencies and skills that includes evidence of an individual's meeting a criterion in areas designated as important. Graduation for students or promotion for employees may be linked to performance at the exhibition. [*See* Demonstration (Performance), Exhibition.]

Exit Interview: an interview with an individual or group that is completing a program or leaving a job. The purpose of the exit interview is to obtain information that can be used to improve the program for future participants or the operation and management of the employer, as well as to bring an appropriate closure to participation in the program or to employment. [*See* Debriefing Interview, Interview.]

Expectation: the anticipated performance of an individual. Assessment of performance can be affected by the expectations of a teacher, parent, friend, or employer, or by the individual's own expectations. Holding high expectations for individuals can increase their likelihood of performing at higher levels; holding low expectations for individuals can reduce their performance. Sometimes, the opposite can occur. Too high expectations can undermine an individual's confidence, while too low expectations can impel an individual to strive for greater performance.

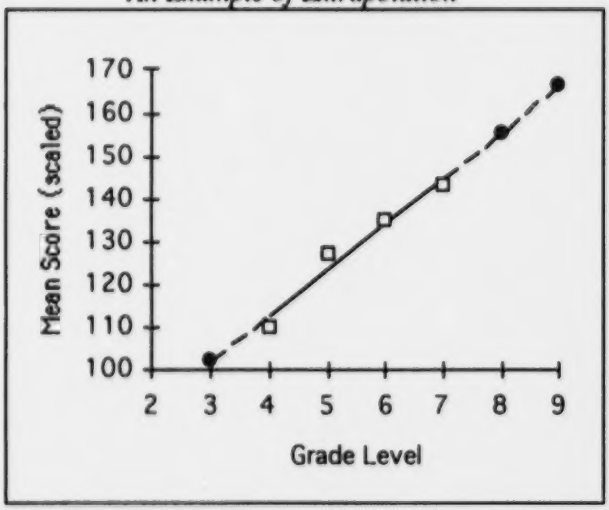
Experience: that which is acquired through exposure to or participation in an activity or process. [*See* On-the-Job Training.]

Expert: one who has demonstrated a high level of proficiency in a knowledge area or skill. [*See* Proficiency.]

Extant Data: existing data, information, and observations that have been collected or that are available for use in the assessment process. [*See* Data.]

Extrapolation: the process of estimating values or scores above and below the range of available data. This is usually done by extending geometric plots (straight lines or curves). Since extrapolations involve assumptions about the nature of performance growth (linear or curvelinear) and judgments in how to extend the graph, and since they are not based on actual data, they should be used with caution. [*See* Grade-Equivalent Score (GE), Interpolation.]

An Example of Extrapolation



▫ = Actual Data Points
--- = Extension of Line
• = Estimations based on extrapolations
Based on extrapolations, the estimated mean scaled scores for the other grades would be 103 for Grade 3, 157 for Grade 8, and 166 for Grade 9.

F

Face Validity: *See* Validity.

Fairness: impartiality. Fairness refers to such aspects of the assessment program as equal opportunity to acquire the knowledge and skills to be assessed, use of developmentally appropriate assessment instruments, use of assessment procedures that are suitable for the individual's characteristics and background, sound administrative conditions, reasonable use of assessment information, and reasonable demands on the individuals being assessed in terms of such factors as time, costs, and personal resources required. [*See* Adverse Impact, Bias, Comparability, Equitable, Equity.]

False Negative: the classification of an individual below the cut score whose true score or typical performance level is actually above the cut level. [*See* Classification Error, Cutting Score, Error of Measurement, False Positive, Passing Score, True Score.]

False Positive: the classification of an individual above the cut score whose true score or typical performance level is actually below the cut level. [*See* Classification Error, Cutting Score, Error of Measurement, False Negative, Passing Score, True Score.]

Feasibility (of Performance Assessment): the likelihood or extent to which appropriate data and information are readily available or can be obtained, produced, or interpreted with available resources such as staff, expertise, time, and equipment. [*See* Teachable.]

Feedback: the information and recommendations provided to an individual about his/her performance based on the results of that individual's assessment and designed to help the individual improve his/her performance and make decisions concerning himself/herself. [*See* Recommendations, Reporting.]

Field Testing: a try-out administration with a sample from the population for which the assessment has been designed and following as closely as possible the administrative conditions for the assessment program when it is fully implemented. The field testing, unlike the pilot testing, usually has a large enough number of participants to examine the implementation's feasibility, and to study the assessment's validity, calculate reliability, conduct item and test analyses, develop norms, and derive related data and information. [*See* Pilot Testing.]

First-Impression Effect: a rater effect in which an assessor or an evaluator tends to base judgments and evaluations on early opinions rather than on a complete picture and tends to distort subsequent information and judgments to fit the initial opinion. [*See* Halo Effect, Rater Effect.]

Floor: the lowest level of performance that can be assessed or measured by an instrument or process. Individuals who are apt to perform near to or below

the floor should be given a less difficult level of the instrument or process in order to obtain a valid estimate of their performance. If the floor of a test is too high for a group of individuals, their distribution of scores will tend to be positively skewed. [*See* Ceiling, Functional-Level Testing, Level (Test), Skewness.]

Form (Test): an alternate version of a test at the same level. If equated, one form may be used in lieu of another when needed to maintain test security or for retesting. [*See* Horizontal Equating, Level (Test).]

Formal: the conducting of an assessment activity in accordance with a prescribed plan, structure, or advance notice. [*See* Assessment Plan, Informal.]

Format: the structure of assessment instruments, evaluation forms, and materials. Format includes shape, size, mode of delivery (e.g., paper or computer), and general design or layout of the materials and equipment.

Formula Score: *See* Correction for Guessing.

Foundation: the rationale for the attributes and domains covered by the assessment. Foundations can include curriculum frameworks, school district goals and objectives, program goals and desired outcomes, business plans, job descriptions, job analyses, governmental regulations and laws, and client and community needs. Through descriptive or prescriptive analysis, foundations become the basis upon which the domains to be covered by the assessment are determined, and the elements, indi-

cators, and descriptors are identified. [*See* Criterion, Model.]

Free-Response Item: *See* Constructed Response Test.

Frequency of Assessment: how often an assessment is administered (e.g., once a year in grades 4 and 8, twice a year in grades 2 and 10; every six months for employees with less than three years of experience, annually for employees with three or more years of experience). [*See* Schedule.]

Functional-Level Testing: the administration of an instrument or process whose difficulty level is appropriate for the individual being tested, but not necessarily for the age or grade level or group membership of that individual. Functional-level testing refers to the individual being tested, not to the group for whom the assessment was designed nor the group to which the individual belongs. For example, if a 7th grader is functioning in the domain being assessed at the same level as many 7th graders and is given a measure designed for 7th graders, then this is functional-level testing for that individual, as well as in-level testing. If this 7th grader is functioning at the level of the typical 5th grader and is given a test designed for 5th graders, then this is functional-level testing for this individual, as well as out-of-level testing. If the latter student who is functioning at a 5th grade level is given a 7th grade test, then this is not functional-level testing. In this case, even though it is in-level testing, the validity and reliability of the results of the test for that individual are questionable as he/she will be performing at or near the floor of that test, and similarly for individuals performing well above

the 7th grade level and who are at or near the ceiling of the test. [*See* Ceiling, Developmentally Appropriate, Floor, In-Level Testing, Level (Test), Out-of-Level Testing, Vertical Equating.]

G

Gain Score: *See* Difference Score.

Generalizability: the appropriateness of using results from one context or purpose in another context or for another purpose. [*See* Context (Learning), Generalizability Theory, Situational Specificity, Transferability, Transportability, Validity, Validity Generalization Study.]

Generalizability Theory: an approach that uses classical test theory's assumptions about error with analysis of variance (ANOVA) techniques to estimate the magnitude of error from various sources. The results of this analysis provide information on the degree of generalizability of the test scores beyond the items, individuals, and settings covered by the analysis. [*See* Classical Test Theory, Error of Measurement, Generalizability.]

Goal: a statement of intent or an end that a person or a group strives to attain. A goal tends to be more general than an objective. For example, a goal could be to learn how to use a new word processing software package. An objective could be to learn how to move text within a document. [*See* Desired Outcome, Instructional Goal, Objectives, Performance Goal, Priority Goal.]

Grade-Equivalent Score (GE): a derived score that represents the median performance of students during each month of the school year. For example, a grade-equivalent score of 3.4 would purport to correspond to the median performance of 3rd graders in the 4th month of the school year. Grade equivalent scores are derived, using interpolation and/or extrapolation procedures, from actual data at one or more points for each grade during the school year. They are based on assumptions about the nature of the learning curve within and across grade levels and over long vacations and summer breaks. In some cases, extrapolations are made well above or below the ceiling or floor of an instrument, thus making scores based on such extrapolations questionable. [*See* Derived Score, Extrapolation, Interpolation.]

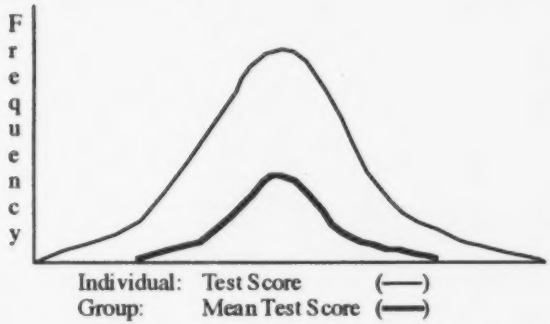
Grading: the process of assigning marks, often letters (A, B, C, D, E) or numerical indices, that signify the quality of an individual's performance on an assessment task or activity.

Grading on the Curve: the assignment of grades to specific assessments or to individuals so that the distributions of grades covers the full range of grades, with most of the grades in the middle of the distribution.

Group Norms: those statistical data that represent the average (usually mean score) results for various groups (e.g., schools, office branches), rather than scores of individuals within one of these groups (local norms) or individuals across all groups (national norms). The variability of average scores,

as in group norms, is less than that of individual scores. [*See* Local Norms, National Norms.]

Distributions of Group and Individual Norms



Group/Collaborative Product: the outcome or result of two or more persons working together to complete an assessment task (e.g., a report based on data collected by four geology students and written as part of a course requirement, a car worked on by three mechanics prior to road testing).

Grouping: a number of individuals classed together for the purpose of: (a) preparing a project to be assessed; (b) collaborating in order to demonstrate competencies; (c) administering a specific difficulty level of a test or assessment; or (d) norming, analyzing, scoring, and/or reporting results on an assessment.

H

Halo Effect: a type of rater effect in which an assessor or an evaluator tends to base overall judgments or evaluations on selected pieces of information rather than on all available relevant information. [*See* First-Impression Effect, Rater Effect.]

Handicapping Condition: a disability (auditory, visual, other physical, emotional, learning) that affects an individual's performance on an assessment if the assessment is administered under standard conditions. Adjustments made for such disabilities include use of an amanuensis, computerized testing, large-print test materials, audiotaped versions of tests for persons with visual disabilities, extended time limits, and special testing locations (e.g., in a hospital). [*See* Amanuensis, Special Testing Arrangements.]

High Inference: the types of items and tasks, judgments, decisions, and conclusions that are based on complex inductive reasoning and that require a high degree of subjectivity on the part of the individual (e.g., because an employee can answer correctly all problems on the test about balancing checkbooks, the supervisor infers that she can keep accurate records of invoices paid and outstanding; since the student is looking over the shoulder of the student in front of him, the proctor assumes that the student is copying the answers; the test taker is asked to identify the sentence that best expresses the writer's

message). [*See* Inference, Low Inference, Subjective.]

High School Equivalency Test: a test that, if passed, is the equivalent of an individual having completed four years of high school in an accredited secondary school. The Tests of General Educational Development (GED Tests) are the most widely used in the United States. Some states have their own tests (e.g., California High School Proficiency Test).

High-Stakes Testing: an assessment to which important consequences, such as graduation or licensure, are attached to the results. The requirement of passage of high-stakes tests is regarded by some as a motivator for individuals to prepare for tests and as a means of giving credibility and value to the purpose of the assessment (e.g., earning a license, receiving special recognition). For others, it is seen as a means of discouraging people from pursuing certain educational or career objectives. [*See* Test (Score) Pollution.]

Higher-Order Thinking Skills (HOTS): those thought processes that are needed to solve problems and make necessary decisions in everyday activities, as well as the mental processes needed to benefit from instruction. Examples of such skills are observing, comparing, classifying, analyzing, evaluating, summarizing, justifying, developing explanations, and making inferences (deductive and inductive). [*See* Cognitive Domain, Inference, Skills Hierarchy, Taxonomy (Cognitive Domain).]

Holistic Scoring: the assignment of a single score that reflects an overall impression of performance on a

measure. Scores are defined by prescribed descriptors of levels of performance, examples of benchmarks at each level, or scoring rubrics. [*See* Analytic Scoring, Benchmark, Primary Trait Scoring, Scoring, Scoring Rubric.]

Horizontal Equating: the process of statistically linking two or more forms of a measure at the same test level onto a common score scale so that the scores or results on one form can be used as estimates of scores or results on the other form(s) to which that form has been horizontally equated. The process involves either two or more randomly equivalent groups each taking a different form, or one group taking two or more parallel forms of a test measuring the same domain or attribute, with the order of the forms taken assigned in a manner that accounts for practice effect (e.g., half take form A first, then form B; the other half take form B first and then form A). Horizontal equating permits retesting with a different form, enhances test security, and, when multiple forms are given to a group at the same time, reduces opportunities to benefit from copying answers. [*See* Anchor, Cheating, Equated Forms, Estimate *(noun)*, Form (Test), Level (Test), Scaled Score, Test Security, Vertical Equating.]

In-Level Testing: the administration of an instrument or process that was designed for the group to which the individuals belong, but may be too easy or too difficult for some of those individuals in the group. For example, an Intermediate level Spanish test, that was designed for persons with a proficiency in Spanish similar to persons who have completed two years of high school Spanish or one year of college Spanish, would be "in-level" for use at the end of a second-year Spanish course in high school or a first-year course in college. However, some individuals in both groups may find the Intermediate level test too difficult and thus need to take the Beginning level test, which others may find it too easy and thus need to take the Advanced level test. In the cases of those individuals who are at or near the floor or ceiling of the Intermediate level test, in-level testing may lead to invalid and unreliable estimates of proficiency in Spanish. In-level testing is frequently done because of administrative feasibility (e.g., same directions, same materials, same timing). However, functional-level testing can lead to more valid and reliable results for all individuals tested. [*See* Ceiling, Floor, Functional-Level Testing, Level (Test), Out-of-Level Testing.]

Incompetence: the intentional or unintentional failure to perform at a minimally acceptable level. Incompetence usually results in remediation, reten-

tion at the current level, or possibly dismissal or transfer. [*See* Remediation.]

Inconsistency: a type of rater effect in which an assessor or an evaluator tends to rate or to interpret similar data and information in different ways. [*See* Consistency, Discrepancy, Rater Effect.]

Indicator: *See* Criterion.

Inference: a logical conclusion or judgment that is explicitly supported by data, evidence, and information gathered as part of the assessment process. [*See* Data, Evidence, High Inference, Higher-Order Thinking Skills (HOTS), Information, Low Inference.]

Informal: the conducting of an assessment activity without a prescribed plan or structure, or with little or no advance notice. [*See* Assessment Plan, Formal.]

Information: the knowledge about the attributes and performance of an individual, based on assessments, documentation, and data sources used in the assessment process.

Informed Consent: the agreement between concerned parties about the data-gathering process, and/or the disclosure, reporting, and/or use of the data, information, and/or results from an individual's assessment. [*See* Confidentiality, Consent, Reporting.]

Instruction: the systematic provision of information, opportunities, and resources to promote the devel-

opment of a repertoire of knowledge and skills. [*See* Training.]

Instructional Goal: a statement of what students are expected to learn in a given lesson, unit, course, program, or across educational and training programs. [*See* Goal, Objectives, Performance Goal.]

Instrument: a device used to collect data, information, and evidence. These devices can include tests, questionnaires, checklists, rating scales, application forms, interview schedules, and observation forms. [*See* Measure (*noun*).]

Intelligence-Fair Testing: based on the work of psychologist Howard Gardner, an approach to testing that considers an individual's skills in terms of seven types of intelligence that are relatively independent of one another. These seven are: logical-mathematical, linguistic, musical, spatial, bodily-kinesthetic, interpersonal, and intrapersonal.

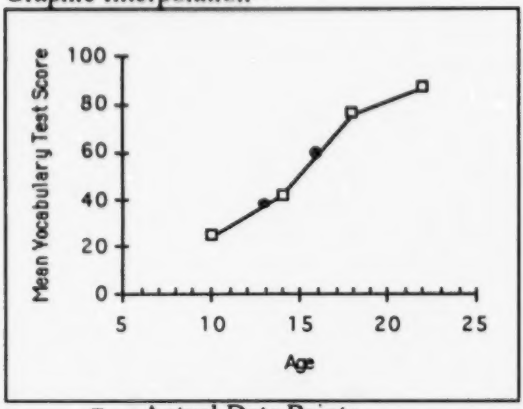
Interdisciplinary: the covering of knowledge and skills from several academic subject areas and/or domains. For example, conducting a science experiment may be considered interdisciplinary if successful completion of it requires written communication skills, specific knowledge in science and mathematics, ability to illustrate results, and skills in using specific equipment.

Internal Consistency: *See* Reliability.

Interpolation: the process of using available points based on actual data to estimate values between two

Two Examples of Interpolation

Graphic Interpolation--



▫ = Actual Data Points
• = Data based on interpolations

Arithmetic Interpolation--

	Actual Mean
Age 10	24
Age 14	42
Age 18	76
Age 22	87

Calculating Interpolated Means for Ages 13 and 16:

$$\text{Age 13} = 24 + \left(\frac{13-10}{14-10}\right)(42-24)$$

$$= 24 + \frac{3}{4}(18) = 24 + 13.5 = 37.5$$

$$\text{Age 16} = 42 + \left(\frac{16-14}{18-14}\right)(76-42)$$

$$= 42 + \frac{1}{2}(34) = 42 + 17 = 59$$

points. Interpolation is usually done by calculation, based on the distance between two known points, though it can also be done geometrically by connecting the points. Interpolation is often based on a linear connection between two points, but in some cases curvelinear procedures are used. [*See* Extrapolation, Grade-Equivalent Score (GE).]

Interview: a series of orally-delivered questions designed to elicit responses concerning attitudes, information, interests, knowledge, and opinions. Interviews may be conducted in person or by telephone, and with one individual or a group. The three major types of interviews are: **(1)** structured (where all questions to be asked by the interviewer are specified in advance); **(2)** semi-structured (where the interviewer can ask other questions and prompts in addition to the specified questions); and **(3)** unstructured (where the interviewer has a list of topics, but no or few specified questions). [*See* Clinical Interviews, Conference, Debriefing Interview, Exit Interview, Prompt, Responses.]

Interviewer: the assessor who conducts the interview, either in a face-to-face setting or by telephone, and makes a record of the responses. [*See* Amanuensis, Assessor.]

Intrinsically Important: that which is essential, significant, or inherently of high value. [*See* Merit, Value.]

Investigation: a systematic examination, observation, or inquiry. Investigations can be: **(1)** a type of assessment task or activity; or **(2)** a process conducted during or after the administration of an assessment

as part of a quality check (e.g., observation of proctors to ensure that they are following procedures and maintaining security of test materials, quality control checks of the scoring process, inquiries into potential cheating cases). [*See* Quality Check.]

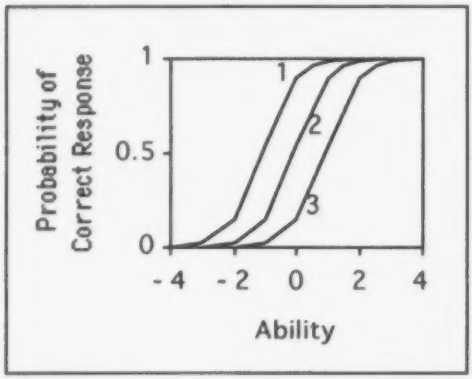
Item: a question on an application form, questionnaire, or test, or in an interview. [*See* Task.]

Item Analysis: the statistical examination of a test item to determine how difficult the item is and how well that item discriminates among test takers who differ on such characteristics as total test score or performance on closely related measures (e.g., course grades, amount of work experience in the area assessed). Item analysis considers such features of the item as the correlation with the total test score (biserial correlation) and with a score on all other items on the test in the same content area (biserial correlation), percentage of all test takers marking an item correctly (item difficulty or p-value), and percentage of test takers in each of four or five groups, based on overall performance level, who marked the item correctly (item discriminating power). [*See* Differential Functioning, Discrimination (Item), p-value, Point-Biserial Correlation, r-Biserial Correlation, Specifications (Assessment), Test Analysis, Validity.]

Item Characteristic Curve (ICC): a graphic representation of the probability of responding correctly to an item based on the overall ability or performance level of the examinee on the attribute being measured by the item (e.g., total test score) and the characteristics of the item itself. There are three logistic models commonly used for item

characteristic curves. The one-parameter model (also called the Rasch Model) considers the probability of a randomly chosen test taker marking the correct answer to an item, based on the ability level of that test taker. The two-parameter model also considers the item's discrimination power. The three-parameter model employs these same two parameters, but also accounts for guessing by low-ability test takers and is sometimes called the pseudochance parameter model. [*See* Item Analysis, Item Response Theory (IRT), Parameter, Rasch Model.]

Item Characteristic Curves for Three Items



Item Difficulty: *See* Item Analysis, p-value.

Item Discriminating Power: *See* Item Analysis.

Item Response: *See* Responses.

Item Response Curve: *See* Item Characteristic Curve (ICC).

Item Response Theory (IRT): an approach to analyzing test items which assumes that the level of performance is related to characteristics of the individuals being assessed as well as to the items themselves. That is, one can estimate the probability of an item being answered correctly, based on relevant information about the individual and the statistical characteristics of the item. [*See* Classical Test Theory, Item Characteristic Curve (ICC).]

Item Specifications: *See* Specifications (Assessment).

Item Type: **(1)** the format or structure of an item (e.g., true-false, multiple-choice, matching, free-response, analogies). **(2)** the content or skill being measured by the item (e.g., math problem solving, map reading, knowledge of plants).

J

Job Analysis: a technique for studying a job in terms of the knowledge, skills, and abilities (KSAs) needed to perform the job in a minimally acceptable manner, and the functions and tasks that are performed as part of the job. The process may involve observation of persons while they are doing their job as well as interviews with these persons, and possibly their supervisors, and review of such documents as job descriptions and progress reports.

Job Description: a summary of the qualifications, duties, responsibilities, physical and mental demands, and working conditions associated with a specific job.

Journal: a daily or weekly record of events which individuals may be asked to keep as part of the instructional activities, jobs, or programs in which they participate. Journal entries may be used to judge writing, progress on projects, and perceptions of experiences. In classrooms, journals can be designed and used to measure changes in writing skills over time. [*See* Log, Student Learning Log.]

Judge: the person who makes judgments for use in assessing individuals and evaluating performance. [*See* Assessor, Evaluator, Judgment, Rater.]

Judgment: an appraisal, decision, or opinion about the performance level of an individual with respect to the knowledge, skill, ability, behavior, or attribute being assessed. [*See* Objective, Subjective.]

K

Key (Scoring or Answer): a listing, or an overlay stencil for an answer sheet, showing the correct responses to a set of test items, usually multiple-choice items. [*See* Scoring Rubric.]

Knowledge: the sum of the information and experience the individual has acquired or learned and is able to recall and use. [*See* Achievement Test, Competency.]

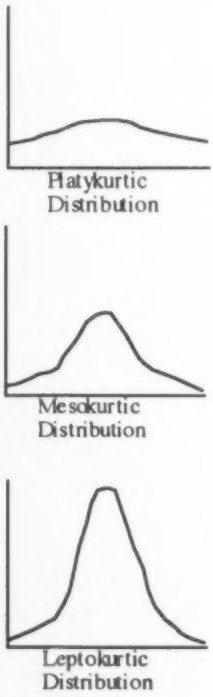
Knowledge-Based Assessment: a process for measuring that focuses on the information and experiences that have been acquired or learned. [*See* Behavior-Based Assessment, Experience, Knowledge, Outcomes-Based Assessment, Performance-Based Assessment, Skills-Based Assessment.]

KSAs: an abbreviation for knowledge, skills, and abilities. [*See* Job Analysis.]

Kuder-Richardson 20 (KR20) Formula: an internal-consistency estimate of reliability, used when items are scored dichotomously (e.g., right/wrong, pass/fail). [*See* Coefficient Alpha, Dichotomous, Reliability.]

Kurtosis: the degree to which a distribution is peaked or flat. A distribution with peakedness similar to that of a normal curve is called mesokurtic. One

that is flat, compared to the normal curve, is a platykurtic distribution and has a higher proportion of cases at the upper and lower ends of the distribution compared to the normal curve. One that is tall and peaked is called leptokurtic, and most cases in the distribution have the same or a very similar score. The degree of kurtosis can be expressed numerically, with negative values associated with platykurtic distributions and positive values with leptokurtic distributions. [*See* Normal Curve, Skewness.]



L

Lake Wobegon Effect: a phenomenon in which most individuals or groups perform above average. This effect was described in John Cannell's 1987 survey report that all of the 50 states and 90% of their school districts reported students had tested above the national average. However, the same states were reporting poor results in terms of graduation and literacy rates. Achievement test scores can produce an exaggerated picture of achievement. Among the causes of the Lake Wobegon Effect are: dated norms, use of nonsecure tests, selection of curricula and programs that are closely aligned to the tests, teaching to the tests, inappropriate test administration and scoring procedures, and selection of those individuals to be tested (i.e., excluding less capable persons). The Lake Wobegon Effect leads to a positive, but misleading, impression of performance. [*See* Test (Score) Pollution.]

Learning Environment: *See* Context (Learning).

Learning Outcomes: the products of instruction or exposure to new knowledge or skills. Examples include mastery of a new skill, successful completion of a course or program, finishing a project or report, or attaining a given level of performance on an assessment. [*See* Outcome Variables, Student Learning Outcomes.]

Leniency: a type of rater effect in which an assessor or an evaluator tends to rate an individual too high or to judge the performance level as better than it actually is. [*See* Rater Effect, Stringency.]

Leptokurtic: *See* Kurtosis.

Lesson: the content that is to be taught or the activity that is to be done during a specific period of instructional time.

Level (Test): a designation of the difficulty for a version of a test that is part of a series of articulated tests. The availability of various levels of tests in terms of difficulty permits functional-level testing, as well as the comparison of performance of individuals and groups at different ages, grades, or proficiency levels. [*See* Articulated Tests, Form (Test), Functional-Level Testing, Vertical Equating.]

Likert Scale: a widely used format, developed by Rensis Likert, for questionnaires and attitude surveys in which individuals are given a continuum of responses from which to chose (e.g., strongly agree, agree, disagree, strongly disagree; frequently, often, occasionally, rarely, never; excellent, good, fair, poor). [*See* Attitude Scale, Scoring Scale.]

Local Norms: those statistical data that portray test performance based on a group of individuals representing the site(s) where the testing is occurring. Local norms are used to interpret performance of an individual in relation to others in that local group. [*See* Group Norms, National Norms.]

Log: a journal or diary, maintained by an individual, that includes such topics as decisions, plans, activities, work undertaken and completed, results, changes, and reflections. The log can serve as a source of information for an assessment or an evaluation, or can be included as part of a portfolio. [*See* Journal, Portfolio, Reflection, Student Learning Log.]

Low Inference: the types of items and tasks, judgments, decisions, and conclusions that require a low degree of subjectivity on the part of the individual (e.g., the librarian determines that an assistant needs more training because of the high number of incorrect reshelvings of books; the manager determines that the types and sizes of errors in counting inventory in the warehouse indicate a need for a change in procedures; the student identifies the spelling errors in an essay). [*See* High Inference, Inference, Objective, Subjective.]

M

Mastery Testing: an assessment process designed to determine whether an individual has acquired a predetermined level of content knowledge or skills. Often, mastery testing is part of an instructional system or training program that involves learning activities and tests which are taken until the learner achieves the pre-established or desired level of mastery and then can move to the next sequenced lesson or instructional unit.

Matrix Sampling: a technique that allows generalizations about mastery of content without the entire test being taken by all individuals (or groups of individuals). Rather than all individuals taking all tests, items, or assessment tasks, each individual (or group) is administered some items or tasks. Collectively, all items and tasks are completed by a number of individuals or groups. This approach reduces the cost and time for administration of the assessments, although production of materials (due to multiple forms) and processing of results may be higher. Matrix sampling provides information about assessments and about groups, but must be used with caution when estimating results for individuals.

Mean: the arithmetic average of a set of data, calculated as the sum of a set of values divided by the number of values. [*See* Average, Central Tendency Measure, Median, Mode, Normal Curve.]

Example of Matrix Sampling

	Group 1	Group 2	Group 3
Writing Mechanics Items	1, 2, 9, 13, 15, ...	3, 5, 6, 10, 14, ...	4, 7, 8, 11, 12, ...
Essay Topic	A, C	C, B	B, A
Documents to Review/ Edit	Letter #11, Progress Report #14	Memo #9, Letter #8	Progress Report #6, Memo #19

Meaningfulness: the value of new content and skills to an individual's existing knowledge. When material is highly relevant to one's prior knowledge and experience, it is more easily acquired and more readily used. Relevant content and skill may reflect an individual's background, culture, ethnicity, and experiences. Content and skills that are of value in an individual's life should be assessed in ways that reflect how the information or skills will be used (e.g., teaching students computation by balancing checkbooks; training food servers in foreign language expressions and vocabulary needed to take restaurant customer orders.)

Measure *(noun)*: an instrument or device that provides data on the quantity or quality of that aspect of performance being assessed. [*See* Assessment, Instrument, Test *(noun)*.]

Measure *(verb)*: to classify or estimate, in relation to a scale, rubric, or standard, the degree of quality or

quantity of that aspect of behavior, learning, or performance being measured. [*See* Assess, Estimate *(noun)*.]

Measurement-Driven Instruction: the type of teaching and learning that is based upon, or heavily influenced by, the assessments used to determine whether the objectives are being met. [*See* Objectives.]

Median: the middle data point or score in a distribution, with half the data above and half below that point. The median is also the 50th percentile. [*See* Average, Central Tendency Measure, Mean, Mode, Normal Curve, Percentile.]

Mentoring: the provision of support by experienced persons to promote the development of new employees or less experienced individuals. Mentors provide support by serving as advisors, coaches, tutors, role models, resources, facilitators, and confidants. [*See* Coaching, Modeling.]

Merit: the overall competence of an individual, as defined by intrinsic qualities and performance levels. Merit can include such factors as breadth and depth of knowledge of an area or topic, specialized training completed, ability to interact with different types of people, special talents, and fluency in a second language. [*See* Competence, Competency, Meritorious Performance, Talent, Value, Worth.]

Meritorious Performance: a level of performance that well exceeds the standard for minimally acceptable and that may be worthy of special recognition, pro-

motion, or reward. [*See* Exemplary Performance, Merit, Minimally Acceptable.]

Mesokurtic: *See* Kurtosis.

Metacognition: the psychological processes that guide individuals through learning tasks. These processes include planning, activating, monitoring, and evaluating of lower-order skills. Self-evaluation of cognitive activities can be developed to enhance performance. [*See* Self-Evaluation.]

Methods of Assessment: the approaches, procedures, or techniques, such as structured interviews or portfolios, that are used to acquire data and information and to conduct the assessment (e.g., for reading skills of 10th grade students, the use of reading logs and written group reports on authors, followed by individual presentations in English class comparing the plots of two short stories; for quality of service of newspaper carriers, a route check to see that newspapers are being bagged, porched, and delivered by 7:00 a.m. plus a telephone survey of every tenth customer concerning satisfaction with the carrier's service.)

Minimally Acceptable: a level of performance that meets the minimum standards, as defined by the criteria (e.g., passing a course, being retained in the current job, receiving a merit badge in Scouting). Any lower level of performance is not acceptable in terms of the purpose of the assessment or evaluation. [*See* Minimally Competent, Passing Score, Performance (Individual), Standard.]

Minimally Competent: the lowest level of competency that is necessary for a specific purpose (e.g., being offered a job, receiving a license). [*See* Competency, Minimally Acceptable, Standard.]

Mode: the data point or score with the greatest frequency of occurrence in a distribution. Since more than one data point could have the same high frequency figure, a distribution can have more than one mode. [*See* Average, Central Tendency Measure, Mean, Median, Normal Curve.]

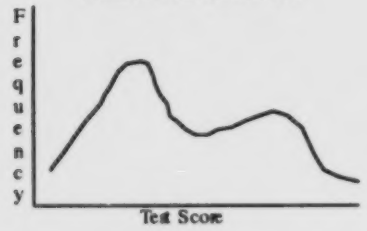
Model: an example of a coherent method, approach, procedure, or strategy of learning or of assessment, as defined by its key or unique assumptions, propositions, attributes, supportive theory, research and practical precedent, or foundation, and which implicitly defines accomplished learning or satisfactory performance. [*See* Criterion, Foundation.]

Modeling: the demonstration or use of exemplary practices, or the exposure to effective behaviors, to promote learning or repertoire expansion in others. [*See* Exemplary Performance, Mentoring.]

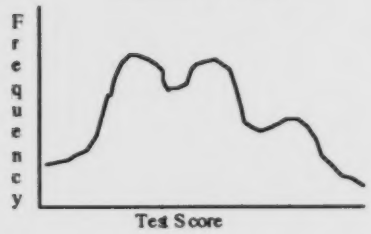
Monitoring: the checking on a process or an individual to verify that progress is being made, required activities are occurring, assessment and evaluation procedures are being implemented, suggested practices are being tried, prior information is still applicable, earlier decisions can still be justified, and/or standards are being met. [*See* Audit.]

Examples of Three Types of Distributions

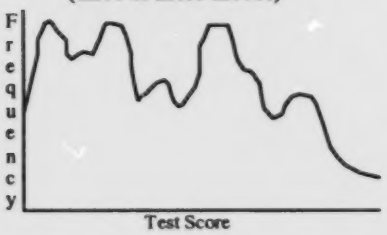
Unimodal Distribution



Bimodal Distribution



Multimodal Distribution
(three or more modes)



Multi-Dimensional: a characteristic of an assessment whereby several measurable aspects of knowledge or skills can be estimated using a single assessment or instrument. An example is a reading test that measures literal comprehension, inferential comprehension, and reading speed. [*See* Scoring Dimension.]

Multiple-Choice Test: a test consisting of items that ask a question or contain an incomplete sentence (i.e., the stem) and present a list of two or more options or alternatives (i.e., the answer and the distractors), from which the test taker selects the best or correct answer(s). Multiple-choice tests are usually in either paper-and pencil or computerized formats, but could be based on actual performance (e.g., asking the child to pick up the red block from among four blocks). [*See* Constructed Response Test, Distractors, Enhanced Multiple-Choice Item, Open-Ended Extended Item, Selected Response Test, Stem.]

Multiple Measures: several different types of evidence that are collected or instruments that are used to better assess an individual's knowledge, skills, and performance. Multiple measures of knowledge of chemistry, for example, might include conducting an experiment on combining several chemical elements, writing an essay on chemical compounds, and taking a paper-and-pencil, multiple-choice achievement test in chemistry. Together, these three measures of chemistry knowledge should provide a more comprehensive, reliable, and valid measure of such knowledge than any one measure alone. [*See* Congruence Analysis, Triangulation.]

N

Narrowing of the Curriculum: the reduction of the content taught to match that represented by the items and tasks in an assessment. [*See* Curriculum Alignment, Teaching to the Test.]

National Assessment of Educational Progress (NAEP): a program, legislated by the U.S. Congress and started in the late sixties, that is used to monitor the achievement of United States students at the elementary, middle, and high school levels. Recently, other types of performance items have been included to supplement the multiple-choice and essay tests.

National Norms: those statistical data that portray test performance of individuals throughout the country, based on a representative sample of those individuals. National norms are used to interpret the performance of individuals in relation to the national sample. [*See* Group Norms, Local Norms, Population, Reference Population, Sample, Sampling of Individuals.]

Negatively Skewed: *See* Ceiling, Skewness.

Norm-Referenced Test (NRT): an instrument developed and used to estimate how the individuals being assessed compare to other individuals in terms of performance on the test. Individual performance is

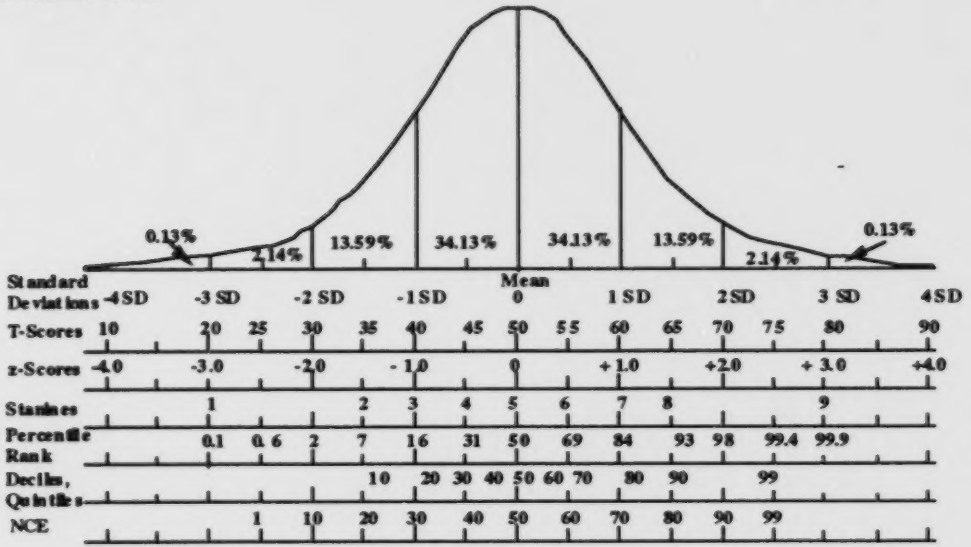
judged in comparison to other individuals tested, rather than against a set of criteria, as with criterion-referenced tests, or to expertise in a broad knowledge area, as with domain-referenced tests. [*See* Criterion-Referenced Test (CRT), Domain-Referenced Test, Local Norms, National Norms.]

Normal Curve: a theoretically derived curve that reflects a distribution of scores or points and that is bell-shaped, symmetrical, unskewed, and unimodal (i.e., has one mode). The mean, median, and mode of a normal curve are identical. The normal curve is frequently used in inferential statistics to determine the probability of a score or point differing from the mean by chance. [*See* Kurtosis, Mean, Median, Mode, Skewness, Standard Deviation (SD).]

Normal Curve Equivalent (NCE): a transformation of the raw scores to a scale with a mean of 50 and a standard deviation of 21.06. NCEs permit conversion of percentile ranks to a scale that has equal intervals of performance differences across the full range of scores and which can be arithmetically manipulated. Percentile ranks cannot be used in arithmetic calculations. (Note: The percentile rank/normal curve equivalent conversion table is provided under "Percentile Rank" in this section of the *Resource Handbook*). [*See* Derived Score, Normal Curve, Percentile Rank (PR), Scaled Score.]

Normative Scale: a series of values that reflects the distribution of scores for a designated group and that permits the expression of an individual's performance results or scores relative to the designated group. [*See* Percentile Rank (PR).]

Normal Curve



BEST COPY AVAILABLE

Norms: those statistical or tabular data on the performance of a designated group on a certain form of an instrument or measure. [*See* Group Norms, Local Norms, National Norms.]

Notes: the descriptive information about the context within which an individual or group is being assessed and/or about the process itself. For a hands-on small appliance repair skills assessment activity, this might include the number and types of individuals present, a list of available supplies and equipment for use in the assessment, a description of the physical arrangements, and information on events that occur during the assessment. [*See* Comments.]

O

Objective *(adjective)*: a characteristic of an assessment, judgment, or conclusion that minimizes the impact of bias and subjectivity, and that yields results which can be empirically verified. [*See* Judgment, Low Inference, Replicable, Subjective, Verification.]

Objectives: the pre-specified intended outcomes of a course, program, process, or policy. In the case of schooling, these are usually in the form of learning and behavioral objectives for students. In work settings, they may be expressed as skills in using new equipment or software, or of acquiring the ability to implement new processes or services, or of marketing certain products and services to new customers. Objectives tend to be more specific than goals. [*See* Benchmark, Desired Outcome, Goal, Instructional Goal, Performance Goal.]

Observation: the careful watching and noting of behaviors and events in order to collect data and information for use in assessing and evaluating an individual's performance. Observations typically occur in the individual's own classroom or work setting, but they may also occur in other settings or off site. Several methods are used to collect data during observations including checklists, coding forms, frequency counts, guided note-taking records, rating scales, and scripting. [*See* Behav-

iors, Context (Learning), Observer, Performance (Individual).]

Observed Score: *See* Obtained Score.

Observer: the person who collects evidence and notes about what he/she is observing. The observer is an assessor, but may or may not be an evaluator. [*See* Assessor, Evaluator, Observation.]

Obtained Score: the actual results for an individual on an assessment administration. Because of error of measurement and other factors, the obtained score may not be the same as that individual's true score. [*See* Error of Measurement, Estimate *(noun)*, Residual Score, True Score.]

On-the-Job Training: the structured provision of training in skills and knowledge through actual work experience rather than indirect methods such as lecturing, role playing, or simulations. [*See* Experience, Mentoring, Training.]

Open-Ended Extended Item: those complex items that can be solved in more than one way and have more than one correct answer. These items require that individuals give reasons for their conclusions. When responding to open-ended extended items, individuals can demonstrate their depth of knowledge and can use a variety of problem-solving approaches. [*See* Constructed Response Test, Enhanced Multiple-Choice Item, Multiple-Choice Test.]

Open-Ended Test: *See* Constructed Response Test.

Operational Definition: a very precise statement about how observed behaviors or events will be interpreted as representing a designated construct. [*See* Construct, Operationalize.]

Operationalize: the defining of a psychological or physical attribute by the way it is measured. For example, knowledge of mathematics is operationalized by the number and types of problems that an individual solves correctly, or an individual's short-term memory is operationalized by the number of digits or letters a person can recall after hearing a list of them presented one time. [*See* Attribute, Construct, Operational Definition.]

Orientation: the provision, to some or all stakeholders, of the performance assessment policies, procedures, and processes in order to promote understanding and to improve the quality of the assessment. [*See* Stakeholders.]

Out-of-Level Testing: the administration of an instrument or process which was designed for a group other than the one to which the individuals who are being tested belong. For example, a reading test designed for 3rd graders, but given to 5th graders, is out-of-level testing. This may be done for administrative convenience (e.g., have a combined classroom with grades 3-5) or to provide functional-level testing for 5th graders who would find the 5th grade test too difficult. Out-of-level testing is appropriate for functional-level testing situations where the in-level test is too easy or too difficult, but can lead to invalid and unreliable estimates of performance when done for administrative convenience. [*See*

Ceiling, Floor, Functional-Level Testing, In-Level Testing, Level (Test).]

Outcome Variables: the results or products of learning. Examples include mastery of skills, completion of courses or training activities, finishing project reports, fulfilling licensure or graduation requirements, and individual performance on standardized tests. [*See* Learning Outcomes, Outcomes-Based Assessment, Outcomes-Based Scoring, Student Learning Outcomes, Variable.]

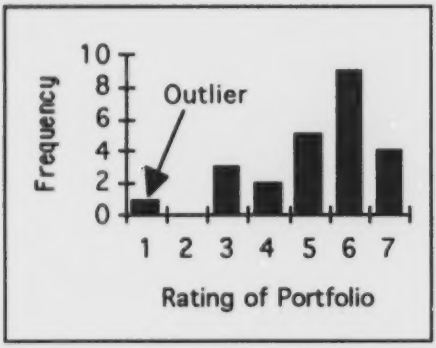
Outcomes-Based Assessment: a process for measuring the extent of attainment of selected outcome variables. [*See* Behavior-Based Assessment, Knowledge-Based Assessment, Outcome Variables, Performance-Based Assessment, Skills-Based Assessment.]

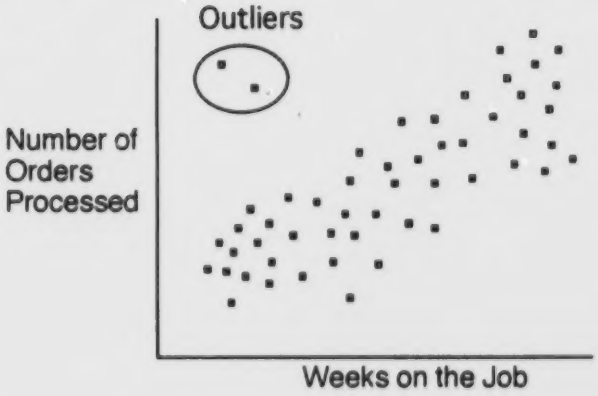
Outcomes-Based Scoring: the determination of an individual's level of performance or degree of proficiency based on the final results or products of an assessment task, rather than on how the task was done. [*See* Outcome Variables, Outcomes-Based Assessment, Performance-Based Scoring.]

Outlier: an extremely high or low score or level of performance in a distribution. The presence of outliers may indicate an invalid test, the administration of an inappropriate test level, incorrect data used, or non-standard test administration practices (e.g., overly hot and uncomfortable testing room, cheating). Outliers should be investigated to determine the reasons behind them and whether the results or data should be used, or if another form or level of the test should be given to, or new data collected

about, the individuals whose scores are outliers in the distribution.

Two Examples Showing Outliers





P

p-value: the percentage, expressed as a decimal, of test takers who responded correctly to an item. The p-value is an estimate of the item's difficulty level. It can have values from zero to 1.00. Items with p-values of .90 or higher are considered to be very easy items for the group of test takers from which the p-values were derived, whereas items with p-values below .30 are quite difficult for this group of test takers. On norm-referenced tests, typically there should be very few items with p-values below .30 or above .90, with the remaining items having p-values spread across the interval between .30 and .90. [*See* Item Analysis, Norm-Referenced Tests.]

Parallel Forms: two or more versions of an assessment that are similar in purposes and content and that have similar statistical characteristics such as means, standard deviations, and correlations with other measures, but that have not been statistically equated to each other. [*See* Alternate Forms, Comparable Forms, Equated Forms.]

Parameter: (1) the actual value of some attribute or variable for a population rather than an estimate of the population value based on the value for a sample within the population. (2) a statistical characteristic of an item (e.g., p-value), as used in item response theory. [*See* Conversion Parameters, Estimate *(noun)*, Item Characteristic Curve (ICC), Item Response Theory (IRT).]

Part Score: the result for a portion of an assessment or a group of items on a test. Also referred to as "subscore." [*See* Scoring Dimension, Subtest, Synthesis.]

Passing Score: a score or level of performance that represents the difference between those individuals whose level of performance is minimally acceptable and those whose performance is not acceptable. Individuals below the "passing score" (also called cut score or critical score) may be enrolled in a special program (e.g., computer-aided instruction center, training course), provided with a tutor, denied a license, or counseled into other courses or jobs. [*See* Coaching, Cutting Score, False Negative, False Positive, Merit, Minimally Acceptable, Remediation, Standard, Standard Setting, Worth.]

Pattern: a series of similar behaviors or common attributes over a period of time or across different settings or contexts.

Peer Review: the assessment and/or evaluation of a student by other students, or an employee by other employees in the same or similar jobs.

Percent: one part of one-hundred units. Percents range from zero to 100. [*See* Percentage.]

Percent Below: the proportion of individuals, expressed as a percent, who performed at a level below the specified score or level.

Percent Correct Score: the number of items that an individual responded to correctly, divided by the total number of items on a test, and expressed as a percent. For example, if a job applicant answered 42 of the 60 items correctly, the percent correct score would be 42 divided by 60, or 70%.

Percent Reaching Standard: the proportion of individuals, expressed as a percent, who performed at or above a specified level (i.e., the standard). [*See* Standard.]

Percentage: the proportion or percent of the total group. [*See* Percent.]

Percentile: one of 99 points that divide a distribution into one-hundred equal parts. For example, a percentile of 42 has 42% of the distribution below it and 58% above it. Percentiles range from 1 to 99. [*See* Decile, Normal Curve, Percentile Rank (PR), Quartile, Quintile.]

Percentile Band: the band defined by the percentile rank corresponding to the score one (or more) standard error below the obtained score and the percentile rank corresponding to the score one (or more) standard error above the obtained score. The percentile rank for an obtained score falls between the lower and upper point of the percentile band, but not necessarily halfway between them. [*See* Confidence Interval, Error of Measurement, Percentile Rank (PR), Standard Error of Measurement (SEM).]

PERCENTILE RANK/NORMAL CURVE EQUIVALENT CONVERSION TABLE

Percentile	NCE	Percentile	NCE	Percentile	NCE
1	1.0	34	41.3	67	59.3
2	6.7	35	41.9	68	59.9
3	10.4	36	42.5	69	60.4
4	13.1	37	43.0	70	61.0
5	15.4	38	43.6	71	61.7
6	17.3	39	44.1	72	62.3
7	18.9	40	44.7	73	62.9
8	20.4	41	45.2	74	63.5
9	21.8	42	45.8	75	64.2
10	23.0	43	46.3	76	64.9
11	24.2	44	46.8	77	65.6
12	25.3	45	47.4	78	66.3
13	26.3	46	47.9	79	67.0
14	27.2	47	48.4	80	67.7
15	28.2	48	48.9	81	68.5
16	29.1	49	49.5	82	69.3
17	29.9	50	50.0	83	70.1
18	30.7	51	50.5	84	70.9
19	31.5	52	51.1	85	71.8
20	32.3	53	51.6	86	72.8
21	33.0	54	52.1	87	73.7
22	33.7	55	52.6	88	74.7
23	34.4	56	53.2	89	75.8
24	35.1	57	53.7	90	77.0
25	35.8	58	54.2	91	78.2
26	36.5	59	54.8	92	79.6
27	37.1	60	55.3	93	81.1
28	37.7	61	55.9	94	82.7
29	38.3	62	56.4	95	84.6
30	39.0	63	57.0	96	86.9
31	39.6	64	57.5	97	89.6
32	40.1	65	58.1	98	93.3
33	40.7	66	58.7	99	99.0

Percentile Rank (PR): the expression of an individual's performance level or score in relation to its rank, or standing, in the distribution of scores of a representative group of individuals. If an individual's percentile rank is 56, for example, then that individual did as well or better than 56% of the group, and worse than 44% of the group. Percentile ranks cannot be arithmetically manipulated due to the varying interval nature of percentile ranks. However, normal curve equivalents can be arithmetically manipulated. (Note: The percentile rank/normal curve equivalent conversion table is provided on the next page.) [*See* Equipercentile Equating, Normal Curve, Normal Curve Equivalent (NCE), Normative Scale, Percentile, Percentile Band.]

Performance (Individual): that which an individual does. Performance depends upon the individual's competence, abilities, and talents as well as upon the context within which the individual learns or works. [*See* Ability, Competence, Context (Learning).]

Performance Appraisal: the systematic process of determining the merit, value, and worth of an individual's current performance and estimating his/her potential level of performance with further development. [*See* Evaluation, Merit, Performance Assessment, Performance Evaluation, Value, Worth.]

Performance Assessment: **(1)** the process of measuring or describing performance attributes of an individual or a group. **(2)** a measurement approach in

which the individual displays behaviors or produces products that are judged by an assessor according to pre-specified standards or scoring rubrics. [*See* Assessor, Attribute, Merit, Performance (Individual), Performance Appraisal, Scoring Rubric, Spontaneous Performance Assessment, Standard, Structured Performance Assessment, Value, Worth.]

Performance-Based Assessment: a measurement approach in which individuals must perform tasks rather than just provide written or oral answers to test questions. Performance is observed and judged against pre-established criteria. For example, a performance-based assessment would require an individual to build a table rather than asking that individual a series of questions on such topics as types of wood, how to measure, and safe use of tools and paints. [*See* Alternative Assessment, Authentic Assessment, Behavior-Based Assessment, Knowledge-Based Assessment, Outcomes-Based Assessment, Performance (Individual), Skills-Based Assessment.]

Performance-Based Scoring: the determination of an individual's level of performance or degree of proficiency, based on the process of doing an assessment task rather than on the final results or products of the task. [*See* Outcomes-Based Scoring, Performance-Based Assessment.]

Performance Domain: *See* Criterion.

Performance Evaluation: the process of determining the merit, value, and worth, based on assessment results, of some performance attribute(s) of the

individual being evaluated. [*See* Attribute, Merit, Performance Appraisal, Value, Worth.]

Performance Goal: a specific statement of what is to be accomplished by the individual (e.g., growth in knowledge, development of a skill, change in behavior), how the goal will be met (e.g., activities, resources, training), when the goal will be met, and how attainment of the goal can be assessed or determined. [*See* Goal, Instructional Goal, Objectives.]

Performance Indicator: *See* Criterion.

Performance Review: *See* Performance Appraisal.

Performance Standard: *See* Standard.

Pilot Testing: a preliminary try-out of a new or revised assessment or process. Pilot testing considers such areas as comprehensiveness and clarity of directions, format of assessment materials, adequacy of assessor training program, and timing of assessment tasks. Since pilot testing is often done with a small sample, it is rarely used to derive statistical data or to obtain an array of benchmarks for scoring. [*See* Field Testing.]

Platykurtic: *See* Kurtosis.

Point-Biserial Correlation: an estimate of the degree of correlation between two variables, one of which is dichotomous (i.e., has only two possible values) and the other continuous. In multiple-choice testing, this is used to correlate performance on an item

(right/wrong) with total test score, as an estimate of the item's ability to distinguish higher scoring test takers from lower scoring individuals and to see if the item and the test are measuring the same or a closely related attribute. [*See* Correlation, Dichotomous, Item Analysis, r-Biserial Correlation, Variable.]

Policy (Assessment): a set of mandates, rules, and guidelines issued by a governmental or administrative agency regarding the purpose of the assessment program and the manner in which it should be conducted (e.g., who will be tested, how often they will be tested, who will administer the assessments, who is authorized to receive results, what information will be given to the test taker, how results may and may not be used, repeated testing, special testing arrangement permitted, handling of irregularities or cheating). [*See* Practice (Assessment), Procedures (Assessment).]

Pool: a collection of items or tasks in a given knowledge or skill area and from which an assessment can be compiled.

Population: the total or entire group of individuals, cases, or events for which an assessment has been designed. Samples are selected from the population for pilot and field testing, for norming and deriving statistical data, and for obtaining benchmarks for use in scoring. [*See* Population Subgroup, Reference Population, Sample, Standardization Sample.]

Population Subgroup: a portion of the total group that has a common characteristics or attribute. For individuals, these characteristics could be defined by

such variables as gender, ethnicity, geographic location, years of work experience, age, and education level completed. [*See* Population, Sample.]

Portfolio: a purposeful collection of documents concerning an individual's performance (e.g., awards, assessment results) and of products produced by the person (e.g., a book report, a drawing of an object, a tape of a talk given to a group, a collection of photographs on a specific topic). The types of documents to be included may be specified, or the individuals may be free to choose what types of documents to include. [*See* Log.]

Positively Skewed: *See* Floor, Skewness.

Power Test: an assessment on which most test takers are able to consider all questions or tasks within the designated time limits. A power test is designed to estimate the maximum performance level, in terms of difficulty level or extent of knowledge and skills, rather than how quickly a test taker can complete a set of items or tasks. [*See* Speed Test, Speededness.]

Practice (Assessment): the manner in which assessments are actually administered and interpreted, whether or not the practice is in accordance with the policy and/or follows the procedures and directions. [*See* Policy (Assessment), Procedures (Assessment).]

Practice Effect: the impact on test performance that results from repeated exposure to similar items, tasks, or testing formats, or to the test-taking

environment and procedures, but not other factors (e.g., additional study, more sleep, quieter testing conditions). The amount of increase in performance tends to diminish over time as individuals become more familiar with various test formats and procedures (e.g., marking scannable answer sheets, keying in responses on a computer, compiling a portfolio). [*See* Test Wiseness.]

Practice Test: a trial assessment completed by the individuals prior to the actual assessment administration and used to familiarize them with the directions, format, procedures, content, and timing. [*See* Coaching.]

Precision: the size of the interval within which a value can be estimated to fall, given a specified level of probability. The smaller the standard deviation becomes with repeated assessment of an individual, given no change in performance that would affect his/her true score, and the smaller the error of measurement, the smaller the size of the interval and the higher the level of precision of an instrument. [*See* Accuracy, Confidence Interval, Consistency, Error of Measurement, Estimate *(noun)*, Obtained Score, Reliability, Residual Score, Standard Deviation (SD), Standard Error of Estimate (SEE), Standard Error of Measurement (SEM), True Score.]

Predictive Validity: *See* Validity.

Preponderance: the emphasis or weight given to the data and information on an attribute of an individual, based on such considerations as the quantity and frequency of occurrence of a behavior, the importance of the attribute to the process of learning

and improving performance, and the potential impact of a behavior or characteristic on other individuals. [*See* Attribute, Weighted Score.]

Prerequisite Knowledge: the prior knowledge that is necessary in order to learn how to solve problems or acquire new skills. For example, knowledge of subtraction is required when learning how to do long division. Familiarity with the periodic table of chemical elements is essential when calculating gas pressure. Knowledge of the meaning of road signs is needed before learning how to safely operate a car on the highway. Understanding how to read blueprints and architectural drawings is needed in order to construct a scale model of a building.

Primary Standards: those standards that apply to the assessment process and instruments rather than to individuals being tested or their performance levels, and which should be met or addressed before the assessment processes are used and the instruments are administered. [*See* Secondary Standards, Standards (Legal), Standards (Professional), Standards (Technical).]

Primary Trait Scoring: the assignment of scores to one or more designated attributes of each task. For example, a set of book reports might be scored for completeness of plot summary, and/or sufficient arguments to support perceived importance of two characters, and/or grammatical accuracy. A set of architectural plans and drawings might be reviewed for accessibility for disabled persons, and/or the use of natural lighting, and/or heating efficiency. [*See* Analytic Scoring, Holistic Scoring, Scoring, Scoring Rubric.]

Priority Goal: (1) among the most important aims that an individual or group is trying to achieve. (2) an important aim or purpose of instruction. For example, in a course on mathematical applications, an individual must be able to interpret charts, tables, and graphs, and to perform basic arithmetic operations. [*See* Goal.]

Procedure-Based Scoring: *See* Performance-Based Scoring.

Procedures (Assessment): the directions for implementing all aspects of the assessment process in accordance with the rules and guidelines given in a governmental or agency policy. Procedures specify how the assessment is to be conducted, designated timelines, persons responsible, forms to be used, documentation to be provided, the data analysis plan, and the steps to be followed. [*See* Data Analysis Plan, Data Collection Procedures, Policy (Assessment), Practice (Assessment), Standardization.]

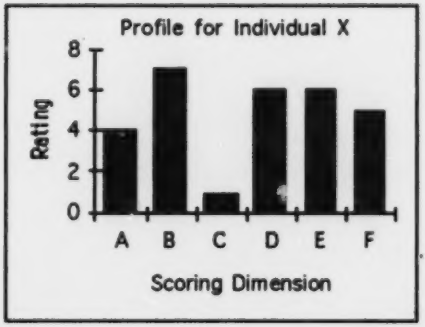
Proctor: a person responsible for the administration of an assessment. Responsibilities of a proctor include distribution and collection of materials and equipment, providing directions, monitoring examinee behavior during the assessment, and ensuring a comfortable testing environment.

Proficiency: sufficient expertise in a knowledge area or adequate mastery of a skill with regard to a standard. [*See* Capacity, Standard.]

Proficiency Test: an instrument designed to determine whether an individual has attained a specified level of expertise or mastery and thus can be promoted or graduated, or considered as having attained a given level of expertise or mastery. [*See* Competency Test, Proficiency.]

Profile: a representation of an individual's performance, or the mean performance of a group, on a number of attributes, measures, or dimensions that use the same scale.

Example of a Profile



Project: a form of complex performance assessment involving several types of activities and products for completion. Most projects involve planning and usually end with a report (written or oral) or product. Examples are: completing several data collection activities on neighborhood needs and writing a report; or planning an experiment on soil erosion and setting up several situations to compare results;

or designing a piece of furniture and building a prototype.

Prompt *(noun)*: a verbal statement or question that provides a cue, reminder, or inspiration, or that motivates the individual being assessed into action. Examples include essay topic questions and follow-up interview questions.

Propriety: the quality of being implemented in an appropriate, legal, and ethical manner; in accordance with specified procedures; and within the policies and guidelines designed to protect the welfare of all affected individuals.

Protocol: (1) the rules and formalities that guide the administration and scoring of an assessment. (2) a record or document of evidence and information related to an assessment. [*See* Scoring Rubric.]

Psychomotor Domain: the range of locomotor behaviors needed to function in the environment and perform tasks, as well as the sensory-motor activities that are essential to learning and communication. [*See* Affective Domain, Cognitive Domain.]

Purpose (of an Assessment): the primary reason or intended use that provides direction for the design, interpretation, and use of an assessment system.

Q

Qualitative Information: the facts and evidence that describe an individual's performance or some phenomenon being assessed, and that typically are recorded in written, audio, or visual form. Statistical and numerical data are not considered qualitative. [*See* Quantitative Information.]

Quality Check: the process of verifying both the accuracy of specific data and information, and the appropriateness of the techniques used to collect, score, or rate, and to analyze the data and information (e.g., trained assessors used, proper timing given for a performance task, right scoring key used for a test, correct formula used to weight various scores). [*See* Accuracy, Audit, Data Analysis Plan, Investigation, Verification.]

Quantitative Information: the facts and evidence that describe an individual's performance or some phenomenon being assessed, and that typically are expressed in numeric, statistical, or graphic form, or can be meaningfully represented by numbers. [*See* Qualitative Information.]

Quartile: one of three points that divide a distribution into four equal parts. For example, the 1st quartile (Q1) has one quarter of the distribution below it and three quarters above it and corresponds to the 25th percentile. The 2nd quartile (Q2) corresponds to the

50th percentile, and the 3rd (Q3) to the 75th percentile. [*See* Normal Curve, Percentile.]

Questionnaire: an instrument consisting of a series of queries and statements that is used to collect data and information from an individual concerning such factors as family background, educational plans and career goals, attitudes and opinions, and outside activities, as well as from others (e.g., teachers, supervisors, parents) concerning the individual's performance. The queries and statements are usually in the form of checklists, multiple-choice items, constructed response items, rankings, ratings, or a combination of these.

Quintile: one of four points that divide a distribution into five equal parts. For example, the 1st quintile has one fifth of the distribution below it and four fifths above it and corresponds to the 20th percentile. The 2nd quintile corresponds to the 40th percentile, the 3rd quintile to the 60th percentile, and the 4th to the 80th percentile. [*See* Normal Curve, Percentile.]

R

r-Biserial Correlation: an estimation of the degree of correlation between two variables, one of which is a continuous variable that has been collapsed into a dichotomous variable and the other is a continuous variable. For example, rather than correlating test scores with the number of invoices prepared correctly, one correlates pass/fail status on the test with the number of invoices prepared correctly. [*See* Correlation, Dichotomous, Point-Biserial Correlation, Variable.]

Range: the difference between the lowest and highest data points, scores, or levels of performance on an assessment for a group or an individual over multiple administrations of the same or equated forms of an assessment. [*See* Restriction of Range, Variability.]

Range Finders: a set of benchmarks consisting of one or more representative examples or typical responses for each scoring point or performance level. [*See* Benchmark.]

Ranking: the process of ordering or arraying from highest to lowest. An individual's level of performance is compared to others in the group, rather than being judged independently of how others in the group performed on the assessment, as is the case with ratings. [*See* Rating.]

Rasch Model: the one-parameter model form of item response theory, developed by Danish mathematician Georg Rasch. [*See* Item Characteristic Curve, Item Response Theory (IRT).]

Rate *(verb)*: to assign judgments or estimations to the magnitude or degree of some aspect of behavior, learning, or performance. [*See* Assess, Rating.]

Rater Drift: the tendency for assessors and evaluators to unintentionally redefine criteria and standards over time or across a series of ratings. [*See* Consistency, Rater Effect.]

Rater Effect: the tendency of an assessor or an evaluator to rate an individual's performance at a level that does not accurately or consistently reflect the level of that individual's performance. There are several types of rater effects, all of which are possible sources of systematic error of measurement. [*See* Assessor, Bias, Central Tendency Effect, Consistency, Contamination, Contrast Effect, Error of Measurement, Evaluator, First-Impression Effect, Halo Effect, Inconsistency, Leniency, Rater Drift, Rating, Reliability, Similar-to-Me Effect, Stringency.]

Rating: a systematic estimation of the magnitude or degree of some behavior or attribute of learning or performance, using a numerical or descriptive continuum. Unlike rankings, ratings can be assigned independently of the performance of others in the group. Ratings are based on scoring rubrics or benchmarks, not in relation to the performance of other individuals. [*See* Behaviorally Anchored

Rating Scale, Benchmark, Range Finders, Ranking, Rater Drift, Rater Effect, Scoring Rubric.]

Raw Score: the first score obtained for an assessment, prior to converting results to scaled scores, percentile ranks, rankings, grades, etc. By itself, a raw score provides little useful information about a person's level of performance. Examples of raw scores include the number right on a multiple-choice test or the initial rating assigned to a portfolio or an essay. [*See* Derived Score, Formula Score, Scaled Score.]

Recommendations: a set of suggestions derived from the assessment results. They may include the use of self-paced instruction to make up for deficiencies, a list of books and articles to read, increased exposure to or more practice with the same learning materials or equipment, alternative techniques to be used in doing one's job, or a list of objectives not met and specific activities to be completed for each one. [*See* Feedback, Reporting.]

Record ***(noun)*:** the written or taped data, evidence, judgments, notes, recommendations, and other statements for use in the assessment process. [*See* Documentation.]

Record ***(verb)*:** to register and store data and other information. [*See* Record ***(noun)***.]

Reductive Reporting Scheme: *See* Reporting Scheme (Reductive).

Reference Population: the total group of individuals for whom an assessment has been designed and from which a norming sample is selected or benchmarks for scoring are chosen. [*See* Population, Standardization Sample.]

Reflection: the process by which an individual reviews his/her past performance as a means of improving future performance. [*See* Log, Metacognition, Self-Evaluation.]

Regression Effect: a tendency for the most extreme scores on the pretesting to move toward the average score of the population or sample on the posttest. Individuals with low pretest scores tend to earn higher scores on the posttest and those with high pretest scores tend to earn lower scores on the posttest. Statistical regression toward the mean is a threat to the validity of the gains that individuals exhibit on a test or assessment. [*See* Difference Score.]

Relative Standard: a cutting score or designated level of performance that is set with reference to how other individuals perform on an assessment and without reference to an absolute performance level. Often the relative standard is used to select the highest or lowest performers in a group when only a certain number can be chosen. For example, the 30 students with the highest scores on a math test will be eligible for the math tournament. Or the 15 qualified applicants who are most proficient in Spanish will be hired as translators for an agency. Or all supervisors with unfair treatment complaints filed by at least five different staff members this past year will participate in the training program.

[*See* Absolute Standard, Classification Error, Cutting Score, Standard.]

Relevance (Domain): the extent to which the domains and indicators covered by a performance assessment system apply to a individual's instructional program, career, or job in terms of both importance and frequency.

Reliability: the degree to which an assessment or instrument consistently measures an attribute. There are several types of reliabilities, for example: (**1**) the degree to which the measure produces consistent results over several administrations assessing the same attribute of an individual (alternate forms, test-retest); (**2**) the degree to which the measure yields consistent results over different administrations with the same individual performing at the same level by the same assessor (intra-rater); (**3**) the degree to which the measure yields similar results for the same individual at the same time with more than one assessor (inter-rater); and (**4**) the degree to which specific observations or items consistently measure the same attribute (internal consistency, coefficient alpha, Kuder-Richardson, split-half reliability). [*See* Coefficient Alpha, Consistency, Error of Measurement, Kuder-Richardson 20 (KR20) Formula, Precision, Reliability Coefficient, Spearman-Brown Prophecy Formula, Split-Half Reliability, Test Analysis.]

Reliability Coefficient: a measure of the reliability of an instrument, ranging from 0.00 to +1.00. The closer the reliability coefficient is to +1.00, the more reliable is the measure. [*See* Reliability.]

Remediation: those techniques or strategies designed to improve an individual's performance in identified deficiencies or specific areas of weakness. [*See* Coaching, Diagnosis, Feedback, Incompetence, Reflection, Self-Evaluation.]

Replicable: an attribute of an assessment, observation system, or research and evaluation study indicating that the process used to obtain the data, scores, observation results, and study findings is explicit and could be repeated. [*See* Objective, Subjective.]

Reporting: the process of communicating results and recommendations to the designated individuals or groups. When reporting to the individual who was assessed, this would be considered part of feedback. [*See* Confidentiality, Consent, Feedback, Informed Consent, Recommendations.]

Reporting Scheme (Complex): a record of evidence describing a performance that provides detailed information about the performance and its context. Scripting of observations can be used in complex reporting schemes. Using the data recorded, the reader can recreate the performance of the person or group observed. For example, activities that occurred are listed; behaviors and responses are recorded; the physical setting is described; the characteristics of the individuals present (e.g., linguistic, gender, ethnicity) are identified. [*See* Reporting Scheme (Reductive).]

Reporting Scheme (Reductive): a record of evidence describing performance that simplifies the data collected through classifying, coding, or analyzing them. For example, an oral report may be described

in terms of separate numerical ratings for accuracy of content, clarity, and persuasiveness. The entire performance could be reduced to three ratings. [*See* Reporting Scheme (Complex).]

Residual Score: (1) the difference between the obtained, or observed, score and the true score that is due to error factors as well as other factors (e.g., test taker's attitude on the day of the test, test taker copying answers from another individual). (2) the difference between the obtained score and a predicted score (e.g., actual college grade-point average versus grade-point average predicted from multiple regression analysis during the college admissions process). [*See* Error of Measurement, Obtained Score, Standard Error of Estimate (SEE), Standard Error of Measurement (SEM), True Score.]

Response Method: the technique by which the individual being assessed indicates responses to items (e.g., giving an oral reply, drawing a chart, keying in the answer on a computer keyboard, pointing at the computer screen, writing an essay, filling in a bubble on an answer sheet, putting a checkmark in front of the correct answer, circling a rating).

Responses: the answers to test, interview, or questionnaire items.

Responsibility: that which a person is expected and obligated to do and for which he/she is accountable. For a student, examples include completion of assigned learning activities and attendance at school. For a word processor, examples are accurate enter-

ing of text, familiarity with the software, and making back-ups regularly.

Restriction of Range: the limited variance of scores for a sample as compared to the variance for the total population from which the sample was drawn. [*See* Population, Range, Sample, Sampling Error, Variability, Variance.]

Review *(verb)*: to examine again or to look at thoroughly in order to assign a grade or rating, make a judgment, come to a conclusion, or evaluate. [*See* Synthesis.]

Rites of Passage: an exhibition that is purposefully designed to be inclusive and comprehensive and that often require public productions. The exhibition must be successfully completed in order for the individual to be promoted or graduated. For example, in order to graduate from high school, a student might be required to exhibit competency in several knowledge domains, complete a portfolio, do a project, and make a presentation before a panel of teachers, other students, and community representatives. To attain a higher rank in Scouting, a youth must earn a number of badges, do several hours of community service, participate in certain activities, and complete other requirements to the satisfaction of the Scout leader as well as a committee of adults from the community. [*See* Exhibition.]

Rubric: *See* Scoring Rubric.

S

Sample: a set or group of individuals, items, tasks, or objects selected from the entire population. [*See* Population, Population Subgroup, Reference Population, Sampling of Content, Sampling of Individuals, Sampling of Performance.]

Sampling Error: the difference between the true value or score for an assessment and the obtained value or score that is due to the limited number of items or tasks on the assessment or of individuals in the standardization sample. Sampling error tends to be smaller for instruments with larger numbers of items and tasks, for assessments with more extensive coverage of the content, and for statistical data based on larger and more heterogeneous samples of individuals. [*See* Error of Measurement, Obtained Score, Reliability, Sampling of Content, Sampling of Individuals, Sampling of Performance, True Score.]

Sampling of Content: the selection of the specific topics and areas of content to be assessed so that the assessment covers the essential content and valid inferences can be made about the amount and type of knowledge that the individual possesses. For example, on a geography test, it is not feasible to cover knowledge of every city in the world; therefore, a sample of questions about 50 cities may be used to estimate an individual's knowledge of world cities. [*See* Breadth and Depth, Error of Meas-

urement, Sample, Sampling Error, Scope of Content, Specifications (Assessment), Validity--Content.]

Sampling of Individuals: the selection of individuals from a population for the norming and the standardization and/or for the selection of benchmarks for use in scoring. To be representative of the population, the sample of individuals should have the same characteristics as the population for variables important to the assessment program (e.g., gender, ethnicity, age, geographic locations, years of experience, licensure status). The size and representativeness of the sample must be large enough to permit valid and reliable inferences to be made about the performance of individuals in the reference population. [*See* Population, Reference Population, Sample, Sampling Error.]

Sampling of Performance: the selection of an array of performances to be assessed that is wide enough in range and large enough in numbers to: (a) cover the scope of the performances addressed by the assessment system in terms of representativeness and comprehensiveness, and (b) permit valid inferences about performances to be made. For example, to determine the physical condition of a firefighter, the individual might be asked to perform a series of activities while in full gear (e.g., run 50 meters, carry a 75-pound person for a distance of 20 meters, climb a 10-meter ladder, drag a fire hose 30 meters) rather than being asked to do every possible physical task that might have to be performed as part of the job. [*See* Error of Measurement, Sample, Sampling Error, Specifications (Assessment).]

Scale (Numerical): a series of numbers or points on a continuum used to assess or judge specific levels of performance. [*See* Scoring Scale.]

Scale (Qualitative): a series of adjectives on a continuum used to assess or judge specific levels of performance. [*See* Scoring Scale.]

Scale (Rating): a series of numerical or descriptive ratings on a continuum used to assess or judge specific levels of performance. [*See* Behaviorally Anchored Rating Scale, Rating, Scoring Scale.]

Scaled Score: a unit based on the transformation of raw scores so that: they can be interpreted in terms of the normal curve, they can be manipulated arithmetically, and the intervals between any two scaled score points can be interpreted in relation to a reference population (e.g., a mean of 10 and a standard deviation of 3, or a mean of 500 and a standard deviation of 100). [*See* Derived Score, Normal Curve, Normal Curve Equivalent (NCE), Raw Score, T-score, z-score.]

Scatterplot: a two-dimensional graph, with each dimension representing a different variable, that contains one point for each individual or case, placed at the intersection of the paired values for the two variables. It is also called a scatter diagram or a scattergram. [*See* Correlation, Variable.]

Schedule: the designated dates and times for various activities related to an assessment (e.g., when the group will do their activity, when the portfolio is to be submitted, when the student will play a musical

piece, when the employees will submit their completed self-assessment questionnaires). The schedule, which is part of the procedures, should also indicate the location of such activities and the people who will be involved. [*See* Assessment Plan, Frequency of Assessment, Timeline.]

School Report Card: a profile or overview of a school's performance made available to the public, typically in a printed document. Often school report cards contain statistical and narrative information describing the school district, the school, and its students. Other types of information included in school report cards are: district demographics, financial data, school enrollment trends, surveys of graduating students' employment and educational plans, dropout rates, student outcomes, school course offerings, types of staffing, and achievements of staff and students in the prior year. The section on student outcomes typically reports standardized achievement test results at selected grade levels. Results of college admissions testing (e.g., SAT, ACT), special programs (e.g., compensatory education, gifted students), and performance assessments may also be reported. The average scores for a school are sometimes compared to state or district averages.

Scope of Content: the extent of coverage, by an instrument or process, of all knowledge, skills, abilities, and behaviors that it measures, in terms of both breadth and depth of coverage. [*See* Breadth and Depth, Sampling of Content.]

Score ***(noun)*****:** the number of points earned on a measure or the degree of success on an assessment of

attributes or performance. Scores are usually expressed in numerical terms, but sometimes in descriptive terms or graphically. [*See* Derived Score, Estimate *(noun)*, Obtained Score, Raw Score.]

Scorer: an assessor who assigns scores and may convert them to derived scores and/or summarize the results of an assessment for use by an evaluator or decision maker. [*See* Assessor, Derived Score, Evaluator, Scoring.]

Scoring: the process of determining the scores and summarizing assessment results. [*See* Analytic Scoring, Holistic Scoring, Primary Trait Scoring, Score, Scorer.]

Scoring Dimension: an attribute or facet of behavior or performance in a domain. For example, in the domain of foreign language proficiency, scoring dimensions could include auditory comprehension, reading, speaking, translating, and writing. In the domain of woodworking, dimensions could include hand tool proficiency, ability to interpret shop drawings, layout proficiency, and knowledge of machine safety. Dimensions are usually determined through logical or statistical analysis, and sometimes are reported as part scores. [*See* Criterion--Domain, Multi-Dimensional, Part Score.]

Scoring Rubric: a set of rules, guidelines, or benchmarks at different levels of performance, or prescribed descriptors for use in quantifying measures of attributes and performance. [*See* Analytic Scoring, Benchmark, Holistic Scoring, Key (Scoring or Answer), Primary Trait Scoring, Protocol, Range Finders, Rating.]

Scoring Scale: a system of ordered values or points at fixed intervals that range from the lowest to the highest possible level of performance on an assessment. [*See* Likert Scale, Scale (Numerical), Scale (Qualitative), Scale (Rating), Score.]

Scripting: the writing of evidence and notes throughout an assessment activity about what is happening (e.g., observing the activities of the botanists on a field trip, watching a paramedic treat an injury, describing the roles and involvement of employees in a brainstorming session, monitoring the development of students' physical skills during the games in basketball season). Scripting does not include making judgments or interpreting the evidence and notes. Scripting may be continuous throughout the activity or intermittent at pre-specified time intervals (e.g., 3 minutes on, 2 minutes off). [*See* Evidence, Notes.]

Secondary Standards: those standards that apply to the assessment process and instruments rather than to individuals being tested or their performance levels, and for which it is desirable, but not crucial or important, that they be met before the assessment processes are used and the instruments are administered. [*See* Primary Standards, Standards (Legal), Standards (Professional), Standards (Technical).]

Selected Response Test: an assessment in which an individual has a set of answers or results from which to choose. [*See* Constructed Response Test, Distractors, Multiple-Choice Test, Stem.]

Self-Evaluation: the process of judging one's own performance for the purpose of self improvement. A person may use such techniques as self-viewing on a videotape, observing and modeling other individuals, self-rating forms, open-ended self-reports, portfolios, or self-study handbooks and materials. [*See* Reflection, Self-Report Measures.]

Self-Report Measures: those instruments in which individuals record their own recollections of events, feelings, judgments, and attitudes. The data recorded using such measures are not the product of ratings or observations by other individuals; thus, they can be biased so as to reflect positively on the respondent. Examples of self-report measures are: teachers asking students to report either orally or in writing on their progress, counselors asking individuals in a smoking-cessation program to report on the number of cigarettes that they have smoked in the past month, or managers asking employees for production data or progress reports. [*See* Self-Evaluation.]

Sensitivity: the awareness and understanding of other people's feelings, attitudes, social and cultural backgrounds, ethnic traditions and customs, languages, interests, rights, and needs.

Setting: the temporal and physical environment of an event or activity. [*See* Context (Learning).]

Short-Answer Item: a constructed response item that requires individuals to generate a brief written or oral answer to a question rather than selecting an answer from a set of options, as in multiple-choice or true-false tests, or generating a longer response,

as in writing an essay or presenting a lengthy oral response. [*See* Constructed Response Test, Item, Multiple-Choice Test.]

Side Effects: *See* Consequences, Unintended Consequences.

Similar-to-Me Effect: a type of rater effect in which an assessor or an evaluator judges more favorably those people seen as similar to himself/herself. [*See* Rater Effect.]

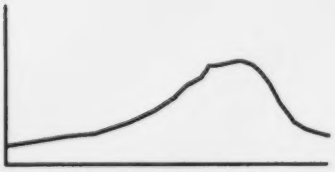
Simulation: an imitation of a typical job task or situation to estimate how well an individual might implement such a task or perform in such an actual situation. Examples are assessing a potential pilot's ability to land a plane safely through a computerized flight trainer program, or having a nurse perform cardiopulmonary resuscitation (CPR) on a training dummy. [*See* Assessment Center, Surrogate Task, Work Simulation.]

Situational Specificity: the extent to which it is appropriate to use an assessment conducted in one setting or context for other settings or contexts. For example, an assessment that is valid for a job or instructional task in one setting might be invalid for the same job or instructional task in a different setting or classroom. [*See* Context (Learning), Generalizability, Transportability, Validity.]

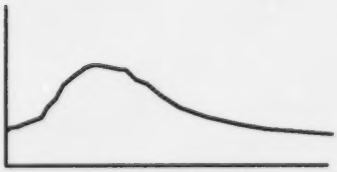
Skewness: the degree to which a distribution is or is not symmetrical. If more than half the data points or scores are below the mean and the distribution's curve tails off to the right, the distribution is posi-

tively skewed. If more than half are above the mean and the distribution's curve tails off to the left, it is negatively skewed. [*See* Ceiling, Floor, Kurtosis, Normal Curve.]

Two Types of Skewness



Negatively Skewed



Positively Skewed

Skill: the ability to use knowledge in a practical manner. [*See* Ability, Achievement Test, Capacity, Competency, Knowledge, Talent.]

Skills-Based Assessment: a process for measuring that focuses on the ability to use knowledge in a practical manner. [*See* Behavior-Based Assessment, Knowledge-Based Assessment, Outcomes-Based Assessment, Performance-Based Assessment, Skill.]

Skills Hierarchy: a concept in which skills are ranked or nested and the learner progresses from simple

skills to more complex skills. In mathematics, progression may be based on a skills hierarchy that extends from counting and computation, to comprehension and application, to analysis of complex problems and theoretical reasoning. Skills hierarchies are often incorporated in content/process matrices which provide an overall summary of the structure of an assessment. [*See* Content/Process Matrix, Higher-Order Thinking Skills (HOTS), Taxonomy (Cognitive Domain).]

Spearman-Brown Prophecy Formula: an equation that permits the estimation of the gain in reliability of a set of test scores if the length of the test were increased with equivalent items. [*See* Reliability, Split-Half Reliability.]

Special Testing Arrangements: the adaptations made to standardized assessment administration conditions to accommodate individuals with special needs (e.g., disabilities, remote work locations). [*See* Amanuensis, Handicapping Condition, Standardized Conditions.]

Specifications (Assessment): a delineation of the major attributes of an assessment (item, task, test) to be developed, including breadth and depth of content to be covered, level of difficulty, format of the assessment materials, supplies and equipment needed, level of complexity, administrative process (e.g., individual or group, location, timing), scoring procedures, and numbers and types of items and tasks. [*See* Breadth and Depth, Content/Process Matrix, Format, Sampling of Content, Sampling of Performance.]

Speed Test: an assessment with a time limit, within which most test takers are unable to complete the set of items or tasks. Examples are measures of reading speed or words-per-minute on a typing test. [*See* Power Test, Speededness.]

Speededness: the extent to which time is a factor in completing an assessment process or instrument. A measure of speededness is the percentage of test takers who complete the assessment or answer all of the test items within the designated time limits. [*See* Power Test, Speed Test.]

Spiraling: the packaging of two or more forms or versions of test booklets in a predetermined order so that when groups of individuals are tested they all do not receive the same form or version. Whenever test booklets are spiraled, they should have common general directions and time limits. [*See* Matrix Sampling, Test Assembly.]

Split-Half Reliability: an estimate of reliability, based on the correlation of the raw score on half of the items with the raw score on the other half of the items (e.g., odd numbered versus even numbered items). This estimate is corrected for test length using the Spearman-Brown Prophecy Formula. [*See* Reliability, Spearman-Brown Prophecy Formula.]

Spontaneous Performance Assessment: a measure based on observing, without prompting, what an individual does during non-assessment activities (e.g., a pediatrician sees the child tie his own shoes at the end of the physical examination, a supervisor overhears an employee responding to a customer's

questions on the store's return policy, a teacher sees a student put on appropriate safety gear prior to working with the equipment in welding class). [*See* Performance Assessment, Prompt, Structured Performance Assessment.]

Stakeholders: those individuals who have a substantial interest in an individual's performance assessment and in the quality of his/her learning and work. These include not only that individual, but also supervisors, teachers, clients, customers, parents, future employers, board members, stockholders, taxpayers, government officials, associations, unions, and community members. [*See* Audience.]

Standard: the level of performance on the criterion being assessed that is considered satisfactory in terms of the purpose of the evaluation. There are three major categories of standards, related to various purposes. **(1)** Developmental standards specify improvement levels to be attained and may be used for planning instructional and on-the-job training programs for individuals. **(2)** Minimum standards designate the level below which performance is not acceptable and are used for such purposes as graduation, hiring, and licensure. **(3)** Desired performance standards reflect what is regarded as accomplished learning or meritorious performance and typically are used for such purposes as promotions, awards, and special recognition. [*See* Absolute Standard, Criterion, Cutting Score, Minimally Acceptable, Passing Score, Relative Standard, Standard Setting.]

Standard Deviation (SD): a statistical estimate of the dispersion of the data values or scores in a distribu-

tion around the mean. For a normal curve, 68.3% of the scores are within one standard deviation above and below the mean, 95.6% within two standard deviations above and below the mean, and 99.7% within three standard deviations above and below the mean. [*See* Mean, Normal Curve, Variability, Variance.]

Standard Error of Estimate (SEE): the standard deviation of the distribution of the differences between obtained scores or actual results and the estimated or predicted values of those scores or results. This statistic is a measure of accuracy or precision of the prediction. [*See* Accuracy, Estimate *(noun)*, Precision, Residual Score, Standard Deviation (SD).]

Standard Error of Measurement (SEM): the standard deviation of the distribution of the errors of measurement (i.e., the difference between the obtained score and the true score, which is known as the error score). [*See* Confidence Interval, Error of Measurement, Residual Score, Standard Deviation (SD).]

Standard Setting: the determination of the individual performance level considered acceptable in terms of the purpose of the assessment program. Standards are usually determined using empirical or judgmental techniques or a combination of these. [*See* Cutting Score, Passing Score, Standard.]

Standardization: the use of consistent procedures for administering, scoring, reviewing, interpreting, and reporting the results of the assessment. [*See*

Procedures (Assessment), Reference Population, Standardization Sample, Standardized Conditions.]

Standardization Sample: the group selected from the reference population for use in standardizing and norming an assessment. [*See* Reference Population, Standardization, Standardized Conditions.]

Standardized Conditions: the administration of an assessment process or instrument to all individuals in the identical manner (e.g., same instructions and timing, comparable testing facilities, access to same equipment and materials, use of trained proctors). [*See* Special Testing Arrangements.]

Standards (Legal): those guidelines and requirements related to an assessment that are specified in the law, governmental policies, and regulations, and by court decisions. (Note: See section of the *Resource Handbook* on "Standards for Assessment Programs and Tests.") [*See* Primary Standards, Secondary Standards.]

Standards (Performance): *See* Standard.

Standards (Professional): those guidelines related to an assessment that are specified by the individuals and associations in the career areas affected, directly or indirectly, by the assessment. (Note: See section of the *Resource Handbook* on "Standards for Assessment Programs and Tests.") [*See* Primary Standards, Secondary Standards.]

Standards (Technical): those guidelines related to an assessment that are specified by psychometricians,

statisticians, test publishers, and specialists in the domains covered by the assessment. (Note: See section of the *Resource Handbook* on "Standards for Assessment Programs and Tests.") [*See* Primary Standards, Secondary Standards.]

Stanine: one of nine points in a normal distribution, each with a width of one half standard deviation (except for stanine values of 1 and 9), and with a stanine value of 5 as the mean of the distribution (stanine is short for "standard-nine"). Twenty percent of the points in the distribution have a stanine value of 5; stanines of 4 and 6 each represent 17% of the points; 3 and 7, each 12% of the points; 2 and 8, each 7% of the points; and 1 and 9, each 4% of the points. [*See* Normal Curve, Standard Deviation (SD).]

Stem: the portion of an item that asks the question or poses the problem. The stem usually is in the form of a question or a sentence to be completed, but may also include such materials as reading passages, maps, charts, formulae, tables, graphs, drawings, and pictures. [*See* Distractors, Multiple-Choice Test.]

Stringency: a type of rater effect in which an assessor or an evaluator tends to rate an individual too low or to judge the performance level as poorer than it actually is. Sometimes stringency is referred to as "negative leniency." [*See* Leniency, Rater Effect.]

Structured Performance Assessment: a measure based on the administration of an assessment instrument or task at a specified time and in a designated manner (e.g., asking all salespersons to

submit a marketing report by the 15th of the each month; giving a timed, running test in physical education class every other Monday; administering an oral reading test in the second-grade classrooms during the first weeks of November and April; assigning history students an essay to write and submit every Friday). [*See* Performance Assessment, Performance-Based Assessment, Spontaneous Performance Assessment.]

Student Learning Log: the notebook kept as a record of student instructional and learning activities. Learning logs are sometime used as a surrogate for a more expensive hands-on performance assessment. [*See* Journal, Log, Surrogate Task.]

Student Learning Outcomes: the measures of student achievement of knowledge and skills and of other educational outcomes such as improved student attitudes and behaviors that should have been taught to them. This term covers acquisition, retention, application, transfer, and adaptability of knowledge and skills. [*See* Learning Outcomes, Outcome Variables.]

Subjective *(adjective)*: a characteristic of an assessment, observation, or conclusion that yields results which cannot be empirically verified by another person. [*See* Audit, High Inference, Judgment, Objective, Replicable, Verification.]

Subscore: *See* Part Score.

Subtest: a section of a test or a set of items within a test used to derive a part score or subscore. [*See* Part Score.]

Summary: a concise report encompassing the major results of an assessment or an evaluation of an individual's performance.

Surrogate Task: a substitute or replacement for an actual activity or task, but one that is very similar to the actual task and that uses the same types of skills, behaviors, materials, and equipment. Examples include asking a candidate for a teaching position to present a lesson to a group of ten students, or having a manager go through a specially-designed collection of in-basket materials and make decisions concerning actions to be taken for each piece, or having a painter paint a piece of sheetrock instead of the walls of a room. [*See* Simulation, Work Sample Task, Work Simulation.]

Synthesis: the process of combining data and information from multiple sources, or ratings and judgments on separate scoring dimensions, or scores on several tests or parts of tests, in order to arrive at a conclusion or result. [*See* Analysis, Configural Scoring Rule, Data Integration, Part Score, Review, Scoring Dimension, Weighted Score.]

System (Assessment): a complete approach to the assessment of performance including its purpose, the rules and regulations that apply, the target group to be assessed, the domains to be covered, the procedures and methods to be employed, the instruments to be used, the persons to be involved, and

the types of scores and reports to be provided. [*See* Testing Program.]

Systemic Validity: *See* Validity.

T

T-score: a scaled score transformed from raw scores to obtain a normal distribution with a mean of 50 and a standard deviation of 10. [*See* Derived Score, Mean, Normal Curve, Scaled Score, Standard Deviation (SD), z-score.]

Talent: an unusually high level of proficiency in performing a task or using a skill. Talents can be in the affective, cognitive, and psychomotor domains. [*See* Ability, Affective Domain, Cognitive Domain, Merit, Proficiency, Psychomotor Domain, Skill.]

Task: an assessment activity or assignment to be completed by the individual or group being assessed. [*See* Item.]

Taxonomy (Cognitive Domain): edited by Benjamin S. Bloom et al. in 1956, the taxonomy is comprised of six hierarchically related cognitive levels including: recall, comprehension, application, analysis, synthesis, and evaluation. Educational objectives can be classified in the taxonomy. The taxonomy has affected curriculum development, instructional methods, and assessment design because it brought cognitive outcomes to educators' attention. It has aided educators in the planning of curricula, instructional activities, and assessments so as to achieve a balance between more and less advanced cognitive skills. Thus, assessments should address the full array of cognitive outcomes and not just the

less complex ones. [*See* Cognitive Domain, Higher-Order Thinking Skills (HOTS), Skills Hierarchy.]

Teachable: the practicality or feasibility of providing instruction on a topic, knowledge area, or skill in terms of the readiness of the learners, the expertise of the teachers, and the availability of resources (e.g., equipment, time). [*See* Attainable, Developmentally Appropriate, Feasibility (of Performance Assessment).]

Teacher-Made Test: an assessment designed by a teacher for everyday classroom use or to evaluate student performance at the end of a lesson, unit, or semester.

Teacher-Outcomes Evaluation: an approach to evaluating teachers that is based on student performance, usually standardized achievement test results.

Teaching to the Test: the focusing of instruction on the content and skills measured by the test and preparing individuals for the format and procedures associated with the specific test to be taken. [*See* Coaching, Curriculum Alignment, Narrowing of the Curriculum, Test (Score) Pollution, Test Wiseness.]

Test ***(noun)*****:** an assessment instrument consisting of a sample of items or tasks from a particular domain and that can provide an estimate of performance in that domain. [*See* Instrument, Measure ***(noun)***, Sampling of Content, Sampling of Performance.]

Test *(verb)*: to administer an instrument or implement an assessment process. [*See* Assess, Assessment, Instrument, Measure *(noun)*.]

Test Analysis: a statistical examination of a test that cuts across the items or tasks on the test. A test analysis may consider the distribution of scores, reliability, speededness, validity, and content coverage (breadth and depth). [*See* Item Analysis, Reliability, Sampling of Content, Speededness, Validity.]

Test Anxiety: a feeling of uneasiness or apprehension about an upcoming test that can affect an individual's performance on the test. A moderate degree of anxiety can motivate the individual to prepare for the test and try harder during the testing, often leading to a higher level of performance than would be the case without such preparation and effort. However, test anxiety can be so high as to interfere with the individual's willingness or ability to prepare for the test and to make a serious effort to perform well, usually leading to a lower level of performance for that individual.

Test Assembly: the process of compiling a test form or version, including the arrangement of items in a reasonable order, provision of standard item and test directions, development of scoring keys, review and editing of the test materials, and other steps undertaken prior to the production and printing of the test materials.

Test Format: *See* Format.

Test (Score) Pollution: an inflation of test scores that is the result of practices, usually associated with high-stakes testing. These practices are designed to increase test scores or performance ratings without a corresponding improvement of actual performance on the attributes being assessed. [*See* Cheating, Coaching, High-Stakes Testing, Lake Wobegon Effect, Teaching to the Test, Test Wiseness.]

Test Security: the protection of the test items and assessment tasks from individuals who will or might take the test or assessment in the future and otherwise benefit or profit from prior knowledge of the test contents. [*See* Cheating.]

Test Specifications: *See* Specifications (Assessment).

Test Wiseness: the use of test-taking skills that will enhance test performance so that the test provides a better estimate of what an individual knows and is able to do. Test wiseness skills include pacing, guessing, and willingness to move on. [*See* Cheating, Coaching, Practice Effect, Teaching to the Test, Test (Score) Pollution.]

Testing Program: a comprehensive assessment system, usually operated by a governmental or private agency, for an intended test-taking population (e.g., all 4th graders in the state, all persons applying for a certain occupational license, all applicants to state colleges). The system typically covers development of the assessment instruments and procedures, production of test materials and related documents (e.g., bulletins of information, practice tests, technical reports, interpretative information), registration of individuals, setting up test centers,

training proctors, administration of the assessments, scoring, reporting results, and maintaining security of both the test materials and the results for individuals. [*See* System (Assessment).]

Testlet: a group of about five to fifteen items or tasks, within a longer test or assessment, which differ from other groups in terms of overall level of difficulty. An individual is given one testlet, and performance on that testlet is used to determine if that individual should receive a more difficult or a less difficult testlet next. The use of testlets can shorten the number of items that need to be answered or of tasks that need to be completed by an individual to obtain a reliable score or estimate of performance for that individual. [*See* Adaptive Testing.]

Timeline: a calendar or list of dates showing the assessment program stages and activities, and indicating the dates by which they should be implemented and be completed. [*See* Assessment Plan, Schedule.]

Track Record: a summary of past events and accomplishments related to an individual's performance (e.g., training completed, course grades earned, awards received, attendance record, participation in relevant activities, disciplinary actions).

Training: the provision of instruction and planned activities to facilitate the learning of specific knowledge, skills, abilities, attitudes, and behaviors. [*See* Instruction, Mentoring, On-the-Job Training.]

Transferability: (1) the degree to which the knowledge and skills demonstrated in solving an assessment task can be used in solving other school/work tasks and real-world activities. (2) one of several characteristics used to evaluate assessments, particularly performance assessments, as well as multiple-choice tests. [*See* Generalizability.]

Transportability: the feasibility of extending the use of a policy, instrument, assessment process, or evaluation system across different groups, approaches, activities, schools and agencies, communities and states, etc. [*See* Generalizability, Situational Specificity, Validity Generalization Study.]

Triangulation: the attempt to obtain more valid results about one aspect of performance by using multiple sources of data, multiple methods of collecting data, and/or multiple interpretations of the same data. [*See* Congruence Analysis, Corroborating Evidence, Multiple Measures.]

True Score: a hypothetical score that represents an assessment result which is entirely free of error. Sometimes true score is thought of as the average score of an infinite series of assessments with the same or exactly equivalent instruments, but with no practice effect or change in the person being assessed across the series of assessments. [*See* Error of Measurement, Estimate *(noun)*, Obtained Score, Practice Effect, Residual Score.]

U

Unintended Consequences: any unplanned or unanticipated outcomes that occur as a result of implementing an assessment. For example, after implementing a minimum competency testing program requirement for graduation, a school experienced an increase in the number of dropouts and it is found that the increase in dropouts is due to the testing program rather than to changes in economic or social factors. The increased number of dropouts, the higher number of library books checked out, and the increased enrollment in the community reading center are examples of unintended consequences of the testing program in this example. Another example is route inspections for postal carriers. Since the carrier is trying to justify the amount of time taken to complete a route, we would expect that the time would be the same or slightly higher during the route inspections. A shorter completion time for a carrier on a designated route would be an unintended outcome of the route inspections. [*See* Consequences, Validity--Consequential Basis of.]

Utility (of an Assessment System): the practical value of an assessment system with respect to such factors as time requirements, logistics, resources needed, costs and benefits, and local and state policies.

V

Validation: the process of determining the appropriateness, meaningfulness, and usefulness of a measurement, an instrument, or an assessment process, and of the inferences made from the results of it. [*See* Validity, Validity Coefficient.]

Validity: the extent to which test scores or responses measure the attribute(s) that they were designed to measure and have the impact that is sought. Several types of validation procedures are described below. [*See* Consequences, Generalizability, Test (Score) Pollution, Validation, Validity Coefficient.]

Concurrent: the relationship of one measure to another simultaneous measure or variable assessing the same or a related attribute.

Consequential Basis of Validity: the assemblage of information on the theoretical and value implications of the way that the results of testing are used, and the appraisal of both the potential and actual social consequences of the testing, including side effects.

Construct: the degree of fit of a measure and its interpretation with its underlying explanatory concepts, theoretical rationales, or foundations.

Content: (1) the appropriateness of the domain definition and the sampling of content. (2) the extent of congruence between the scope of a content area that a measurement claims to cover and what it actually does cover. Both definitions are aspects of construct validity.

Criterion-Related: the correlation or extent of agreement of the test score from an assessment with one or more external variables that measure the attribute being assessed.

Curricular: the extent to which the items on the assessment or test measure the content of a local curriculum, or the extent of agreement between the test coverage (topics, breadth and depth, skills, cognitive complexity) and the goals and objectives of the curriculum.

Evidential Basis of Validity: the assemblage of information about the construct validity of the test scores and measurements, as well as the relevance of the measurement to its applied purpose and its utility in the applied setting. Construct validity, in this case, includes information collected from content and criterion-related validation procedures.

Face: the perceived extent of acceptability or legitimacy of an instrument or process to students, employees, managers, policy makers, parents, the general public, and other stakeholders. [*See* Stakeholders.]

Instructional: the degree to which the items on a test measure: (a) what is actually being taught, and (b) what the individuals being assessed have had an opportunity to learn.

Predictive: the relationship of a measure to performance in a future context such as college or employment or to the results obtained on a future measure or variable assessing a similar or a different (but presumably related) attribute.

Systemic: the negative and positive consequences of testing that could be monitored in order to evaluate the long-range value of the test.

Validity Coefficient: a measure of the degree of validity, usually expressed as the correlation between the measure in question and another measure or a variable. The range is from -1.00 to +1.00, with zero indicating no relationship and values further from zero, a stronger relationship (positive or negative) between the measure in question and the other measure or variable. [*See* Correlation, Validation, Validity.]

Validity Generalization Study: an investigation of the degree of generalizability or of transportability of a policy, instrument, process, procedure, or evaluation system. [*See* Generalizability, Transportability.]

Value: an estimation or a measure of the merit and/or worth of an individual in terms of the intrinsic qualities of the individual (merit) and of the

individual's potential benefit to others (worth). [*See* Merit, Worth.]

Variability: the dispersion or spread of data points, scores, or performance levels within a group. [*See* Range, Restriction of Range, Standard Deviation (SD), Variance.]

Variable: a behavior, characteristic, or event that can change in value from one context to another or over time, or from one individual or group to another (e.g., gender, height, proficiency in using a piece of machinery, fluency in a second language). [*See* Criterion, Outcome Variable.]

Variance: a statistical index of the dispersion of a set of scores or data values around the mean for a distribution. It is calculated as the average of the squared differences between each score or value, and the mean score or value. The square root of the variance is the standard deviation. [*See* Mean, Standard Deviation (SD), Variability.]

Verification: the process of checking the accuracy of data and information on an individual (e.g., comparing job application or resume on college degrees earned with official college transcripts, calling former employers to confirm periods of employment listed on the job application). [*See* Audit, Congruence Analysis, Quality Check.]

Vertical Equating: the process of statistically linking two or more levels of a measure onto a common score scale so that the scores or results on one level can be used as estimates of scores or results on the

other level(s) to which that level has been vertically equated. The process involves a group of individuals taking two or more adjacent levels of a test measuring the same domain or attribute, with the order of the levels taken assigned in a manner that accounts for practice effect (e.g., half take level 3 first and then level 4; the other half take level 4 and then level 3). Vertical equating is needed for articulated tests and permits functional-level testing in groups where performance of individual members must be on the same scale. Vertical equating facilities the interpretation of cross-sectional and longitudinal study data. [*See* Anchor, Articulated Tests, Equated Forms, Estimate *(noun)*, Form (Test), Functional-Level Testing, Horizontal Equating, Level (Test), Scaled Score.]

W-Z

Weighted Score: a score adjusted by such factors as the importance of the attribute assessed to performance, or the reliability and validity of the assessment from which the score was derived, or a combination of such factors. [*See* Preponderance, Synthesis.]

Work Sample Test: an assessment method that uses actual and typical on-the-job activities or tasks (e.g., asking a librarian to locate certain information in the reference section; asking a food clerk to total up a customer's purchases). [*See* Surrogate Task, Work Simulation.]

Work Simulation: a surrogate or imitation of a work sample task (e.g., asking the librarian where she would look for certain information in the reference section; asking the food clerk to demonstrate how he operates all the equipment in determining what a customer owes, such as the bar code scanner, scales, produce code list, and cash register). [*See* Assessment Center, Simulation, Surrogate Task, Work Sample Task.]

Worth: the extrinsic value of an individual to other individuals, employers, and the community. For example, being able to speak Spanish could be of high worth in a school that does not have enough Spanish-speaking teachers, whereas being able to coach basketball may be of low worth to a school if

there are already several teachers who can coach basketball. Knowing Spanish and being able to coach basketball still reflect teacher merit and have value, and may or may not have worth at another school or at another time in the same school. [*See* Merit, Value.]

z-score: a scaled score transformed from raw scores to obtain a normal distribution with a mean of 0 and a standard deviation of 1. [*See* Derived Score, Mean, Normal Curve, Scaled Score, Standard Deviation (SD), T-score.]

RESOURCES IN PERFORMANCE ASSESSMENT AND MEASUREMENT

This section of the *Resource Handbook* lists a variety of types of resources available for further information on performance assessment and measurement. They include professional associations and agencies, test publishers, and printed materials (books, articles, sources of information on tests, and other reference materials).

ASSOCIATIONS AND AGENCIES

Listed below are several national associations and agencies that provide materials and/or services related to performance assessment and measurement. Additional resources include: state educational agencies; national, state, and regional associations for various career fields and professions; licensing and credentialing boards and agencies; the regional educational laboratories and national R & D centers funded by the U.S. Department of Education; the departments of psychology and schools of education and of business at colleges and universities; and the test publishers listed in the next part of this section. The inclusion of an agency or association on this list is in no way to be taken as an endorsement of their services and products; users should review such services and products in terms of their own needs.

American College Testing Program (ACT)
2201 North Dodge Street
P. O. Box 168
Iowa City, IA 52243
(319) 337-1000

American Counseling Association (ACA)
5999 Stevenson Avenue
Alexandria, VA 22304
(703) 823-9800

American Educational Research Association (AERA)
1230 Seventeenth Street, N.W.
Washington, DC 20036
(202) 223-9485

American Evaluation Association (AEA)
c/o School of Education
University of North Carolina at Greensboro
Greensboro, NC 27412-5001
(919) 334-4095

American Psychological Association (APA)
750 First Street, N.E.
Washington, DC 20002-4242
(202) 336-5500

American Psychological Society (APS)
1010 Vermont Avenue, N.W.
Suite 1100
Washington, DC 20005-4907
(202) 783-2077

American Society for Training and Development (ASTD)
1640 King Street
P. O. Box 1443
Alexandria, VA 22313-2043
(703) 683-8100

American Speech-Langauge-Hearing Association (ASHA)
10801 Rockville Pike
Rockville, MD 20852
(301) 897-5700

Association for Supervision and Curriculum Development (ASCD)
1250 North Pitt Street
Alexandria, VA 22314-1403
(703) 549-9110

Buros Institute of Mental Measurements
The University of Nebraska--Lincoln
P. O. Box 880348
135 Bancroft Hall
Lincoln, NE 68588-0348
(402) 472-6203

The Center for the Study of Evaluation
Center for Research on Evaluation, Standards, and Student Testing (CRESST)
University of California at Los Angeles
Graduate School of Education
405 Hilgard Avenue
Los Angeles, CA 90024-1521
(310) 206-1532

The College Board
45 Columbus Avenue
New York, NY 10023-6992
(212) 713-8000

Council of Chief State School Officers (CCSSO)
One Massachusetts Avenue, N.W.
Suite 700
Washington, DC 20001-1431
(202) 336-7048

Educational Testing Service (ETS)
Rosedale Road
Princeton, NJ 08541
(609) 921-9000

EREAPA Associates
2840 Waverley Way
Livermore, CA 94550-1740
(510) 443-5526

ERIC Clearinghouse on Assessment and Evaluation
Catholic University of America
209 O'Boyle Hall
Washington, DC 20064-3893
(202) 319-5120

FairTest
National Center for Fair & Open Testing
342 Broadway
Cambridge, MA 02139-1802
(617) 864-4810

Far West Laboratory for Educational Research and Development (FWL)
730 Harrison Street
San Francisco, CA 94107-1242
(415) 565-3000

Mid-continental Regional Educational Laboratory (McREL)
2550 South Parker Road, Suite 500
Aurora, CO 80014
(303) 337-0990

National Assessment of Educational Progress (NAEP)
Educational Testing Service
Princeton, NJ 08541
(609) 921-9000

National Assessment Governing Board (NAGB)
800 North Capitol Street, N.W.
Suite 825
Washington, DC 20002-4233
(202) 357-6938

National Center for the Study of Writing
University of California at Berkeley
5513 Tolman Hall
Berkeley, CA 94720
(510) 643-7022

National Council of Teachers of Mathematics (NCTM)
1906 Association Drive
Reston, VA 22091
(703) 620-9840

National Council on Measurement in Education
(NCME)
1230 Seventh Street, N.W.
Washington, DC 20036
(202) 223-9485

National Society for Performance and Instruction
1126 Sixteenth Street, N.W.
Suite 102
Washington, DC 20036

Northwest Regional Educational Laboratory
(NWREL)
The Test Center
101 S.W. Main Street, Suite 500
Portland, OR 97204-3212
(503) 275-9500

Society for Human Resource Management
(formerly the American Society for Personnel Administration)
606 North Washington Street
Alexandria, VA 22314
(703) 548-3440

Society for Industrial and Organizational Psychology, Inc. (APA Division 14)
657 East Golf Road, Suite 309
Arlington Heights, IL 60005
(708) 640-0068

Western Michigan University
The Evaluation Center
Center for Research on Educational Accountability and Teacher Evaluation (CREATE)
Kalamazoo, MI 49008-5178
(616) 387-5895

MAJOR TEST PUBLISHERS

Below are listed some of the larger test publishers in the United States. In addition to publishing tests, forms, and performance assessments, many also offer books on assessment, research reports, testing equipment, scanners, scoring services, software, computerized testing systems, and/or contractual services for testing programs and technical work.

The inclusion of a publisher on this list is in no way to be taken as an endorsement of their products and services; users should review such products and services in terms of their own assessment needs. Additional information about these and other publishers may be found in the materials listed under "Sources of Information on Tests" in this *Resource Handbook*.

Academic Therapy Publications
20 Commercial Blvd.
Novato, CA 94949-6191
(415) 883-3314; (800) 422-7249
FAX (415) 833-3720

American College Testing Program (ACT)
2201 North Dodge Street
P. O. Box 168
Iowa City, IA 52243
(319) 337-1000
FAX (319) 339-3020

American Guidance Service
4201 Woodland Road
P. O. Box 99
Circle Pines, MN 55014-1796
(800) 328-2560
FAX (612) 786-9077

American Testronics, Inc.
213 East Tenth Street
Coralville, IA 52241
(319) 351-9086
FAX (319) 351-8327

Assessment Systems Corporation
2233 University Avenue, Suite 200
St. Paul, MN 55114
(612) 647-9220
FAX (612) 647-0412

Consulting Psychologists Press, Inc.
3803 East Bayshore Road
P. O. Box 10096
Palo Alto, CA 94303
(415) 969-8901; (800) 624-1765
FAX (415) 969-8608

CTB/McGraw-Hill
Publishers Test Service
20 Ryan Ranch Road
Monterey, CA 93940-5703
(408) 393-0700; (800) 538-9547
FAX (800) 282-0266

Educational and Industrial Testing Service
P. O. Box 7234
San Diego, CA 92167
(619) 222-1666
FAX (619) 226-1666

Educational Records Bureau
P. O. Box 6650
Princeton, NJ 08541-6650
(609) 243-8596

Educational Testing Service
Princeton, NJ 08541-0001
(609) 921-9000
FAX (609) 734-5410

General Educational Development Testing Service
American Council on Education
One DuPont Circle, Suite 20
Washington, DC 20036-1163

Institute for Personality and Ability Testing, Inc.
P. O. Box 1188
Champaign, IL 61824-1188
(217) 352-4739; (800) 225-4728
FAX (217) 352-9674

Jastak Associates
Wide Range, Inc.
P. O. Box 3410
Wilmington, DE 19804-0250
(800) 221-9728
FAX (302) 652-1644

National Association of Secondary School Principals
(NASSP)
Professional Assistance
1904 Association Drive
Reston, VA 22091-1537
(703) 860-0200; (800) 253-7746
FAX (703) 476-5432

National Computer Systems
Assessment Division
P. O. Box 1416
Minneapolis, MN 55440
(612) 939-5151; (800) 627-7271
FAX (612) 939-5199

Personnel Decision, Inc.
2000 Plaza VII
45 South 7th Street
Minneapolis, MN 55402

PRO-ED, Inc.
8700 Shoal Creek Blvd.
Austin, TX 78757-6897
(512) 451-3246
FAX (512) 451-8542

Psychological Assessment Resources, Inc.
P. O. Box 998
Odessa, FL 33556
(813) 968-3003; (800) 331-8378
FAX (800) 727-9329

The Psychological Corporation
Harcourt Brace Jovanovich, Inc.
555 Academic Court
San Antonio, TX 78204-2948
(210) 299-1061; (800) 228-0752 (educational, speech/language/hearing, psychological);
(800) 634-0424 (human resources)
FAX (800) 232-1223

Psychologists and Educators, Inc.
P. O. Box 513
Chesterfield, MO 63006
(314) 536-2366
FAX (314) 878-3090

Psychometric Affiliates
P. O. Box 807
Murfreesboro, TN 37133-0807
(615) 890-6296; (615) 898-2565

The Riverside Publishing Company
(a Houghton Mifflin company)
8420 Bryn Mawr Avenue
Chicago, IL 60631
(312) 714-7000; (800) 323-9540; (800) 767-8378 (clinical)
FAX (312) 693-0325

Scantron Corporation
1361 Valencia Avenue
Tustin, CA 92680-6463
(714) 259-8887; (800) 722-6876
FAX (714) 259-8423

Scholastic Testing Service, Inc.
480 Meyer Road
Bensenville, IL 60106-1617
(708) 766-7150; (800) 642-6787
FAX (708) 766-8054

Slosson Educational Publications, Inc.
P. O. Box 280
East Aurora, NY 14052-0280
(716) 652-0930; (800) 828-4800
FAX (716) 655-3840

SRA/London House
9701 West Higgins Road
Rosemont, IL 60018
(708) 292-1900; (800) 237-7685
FAX (708) 292-3402

Stoelting Company
620 Wheat Lane
Wood Dale, IL 60191
(708) 860-9700
FAX (708) 860-9775

Touchstone Applied Science Associates (TASA), Inc.
Fields Lane, P. O. Box 382
Brewster, NY 10509
(914) 277-4900
FAX (914) 277-3548

United Educational Services, Inc.
P. O. Box 1099
Buffalo, NY 14224-8099
(716) 668-7691; (800) 458-7900
FAX (716) 668-7875

Western Psychological Services
12031 Wilshire Blvd.
Los Angeles, CA 90025-1251
(310) 478-2061; (800) 648-8857
FAX (310) 478-7838

PRINTED MATERIALS

Below are listed books, articles, NCME instructional modules, sources of information on tests, and other reference books related to performance assessment and measurement. This list is not intended to be a comprehensive list; such a list would take up several hundred pages. However, this list provides a variety of types of materials, including some for lay readers and others for those with relatively advanced levels of expertise in psychometrics.

The inclusion of a document on these lists should in no way be taken as a recommendation for its use or our agreement with the statements in that document. Omission of an item from these lists does not mean that an item is not of high quality and potentially useful. Most of the documents listed below contain bibliographies and will direct the user to additional sources of information on specific performance assessment and measurement topics.

Books, Chapters:

Allen, Mary J.; & Yen, Wendy M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Anastasi, Anne. (Ed.). (1966). *Testing problems in perspective*. Washington, DC: American Council on Education.

Anastasi, Anne. (1988). *Psychological testing* (6th ed.). New York, NY: Macmillan Publishing Company, Inc.

Andrich, David. (1988). *Rasch models for measurement*. Beverly Hills, CA: Sage Publications, Inc.

Angelo, Thomas A.; & Cross, K. Patricia. (1993). *Classroom assessment techniques: A handbook for college teachers* (2nd ed.). San Francisco, CA: Jossey-Bass Inc., Publishers.

Angoff, William H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service. Originally in Robert L. Thorndike (Ed.), (1971), *Educational measurement* (2nd ed.) (pp. 508-600). Washington, DC: American Council on Education.

Archbald, D. A.; & Newman, F. M. (1988). *Beyond standardized testing: Authentic academic achievement in the secondary school*. Reston, VA: National Association of Secondary School Principals.

Astin, Alexander. (1991). *Assessment for excellence*. New York, NY: Macmillan.

Baker, Frank B. (1985). *The basics of item response theory*. Portsmouth, NH: Heinemann Educational Publications.

Barrs, M.; & Thomas, A. (1989). *The primary language record handbook*. New York, NY: Heinemann Educational Books.

Bennett, Randy Elliot; & Ward, William C. (Eds.). (1993). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Berk, Ronald A. (Ed.). (1980). *Criterion-referenced measurement: The state of the art*. Baltimore, MD: The Johns Hopkins University Press.

Berk, Ronald A. (Ed.). (1982). *Handbook of methods for detecting test bias*. Baltimore, MD: The Johns Hopkins University Press.

Berk, Ronald A. (Ed.). (1984). *A guide to criterion-referenced test construction*. Baltimore, MD: The Johns Hopkins University Press.

Berk, Ronald A. (Ed.). (1986). *Performance assessment: Methods and applications*. Baltimore, MD: The Johns Hopkins University Press.

Bloom, Benjamin S.; Engelhart, M. D.; Furst, E. J.; Hill, W. H.; & Krathwohl, D. R. (1956). *Taxonomy of educational objectives, Handbook 1: Cognitive domain*. New York, NY: David McKay.

Bloom, Benjamin S.; Hastings, J. T.; & Madaus, George F. (1971). *Handbook on formative and summative evaluation of student learning*. New York, NY: McGraw-Hill.

Brennan, Robert L. (1992). *Elements of generalizability theory* (rev. ed.). Iowa City, IA: ACT Publications.

Brown, Frederick G. (1983). *Principles of educational and psychological testing* (3rd ed.). New York, NY: Holt, Rinehart & Winston.

Bunda, Mary Anne; & Sanders, James R. (Eds.). (1979). *Practice and problems in competency-based measurement*. Washington, DC: National Council on Measurement in Education.

Cannell, John J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all fifty states are above the national average*. Daniels, WV: Friends for Education.

Centra, John A. (1982). *Determining faculty effectiveness: Assessing teaching, research, and service for personnel decisions and improvement.* San Francisco, CA: Jossey-Bass, Inc., Publishers.

Cole, Nancy S.; & Moss, Pamela A. (1989). Bias in test use. In Robert L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 201-219). New York, NY: American Council on Education; & Macmillan Publishing Company.

Crocker, Linda; & Algina, James. (1986). *Introduction to classical and modern test theory.* Orlando, FL: Holt, Rinehart & Winston.

Cronbach, Lee J. (1990). *Essentials of psychological testing* (5th ed.). New York, NY: Harper Collins College.

Cronbach, Lee J.; & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, IL: University of Illinois Press.

Cronbach, Lee J.; Gleser, G. C.; Nanda, H.; & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York, NY: Wiley.

Donovan, Jerome J. (Ed.). (1968). *Recruitment and selection in the public service.* Chicago, IL: International Personnel Management Association.

Dunnette, Marvin D.; & Hough, Leaetta M. (Eds.). (1990-1992). *Handbook of industrial and organizational psychology.* Palo Alto, CA: Consulting Psychologists Press, Inc.

Ebel, Robert L. (1979). *Essentials of educational measurement* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Ebel, Robert L. (1980). *Practical problems in educational measurement*. Lexington, MA: D. C. Heath.

Eder, Robert W.; & Ferris, Gerald R. (Eds.). (1989). *The employment interview: Theory, research, and practice*. Newbury Park, CA: Sage Publications, Inc.

Educational Testing Service. (1987, May). *Learning by doing: A manual for teaching and assessing higher-order thinking skills in science and mathematics* (NAEP Report 17-H05-80). Princeton, NJ: Author.

Embretson, Susan E. (Ed.). (1985). *Test design: Developments in psychology and psychometrics*. Orlando, FL: Academic Press.

Erwin, T. Dary. (1991). *Assessing student learning and development*. San Francisco, CA: Jossey-Bass Inc., Publishers.

Feldt, Leonard S.; & Brennan, Robert L. (1989). Reliability. In Robert L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 105-146). New York, NY: American Council on Education; & Macmillan Publishing Company.

Finch, Fredrick L. (1991). *Educational performance assessment*. Chicago, IL: Riverside Publishing Company.

Fitzpatrick, R.; & Morrison, E. J. (1971). Performance and product evaluation. In Robert L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Flanagan, John C. (1974). *Measuring human performance*. Palo Alto, CA: American Institutes for Research.

Gable, Robert K. (1986). *Instrument development in the affective domain*. Boston, MA: Kluwer.

Goodman, Kenneth S.; Goodman, Y. M.; & Hood, W. J. (Eds.). (1989). *The whole language evaluation book*. Portsmouth, NH: Heinemann Educational Books.

Green, B. F. (Ed.). (1981). *Issues in testing: Coaching, disclosure, and ethnic bias*. San Francisco, CA: Jossey-Bass, Inc., Publishers.

Greenberg, K. L.; Wiener, H. S.; & Donovan, R. A. (Eds.). (1986). *Writing assessment: Issues and strategies*. New York, NY: Longman.

Gronlund, Norman E. (1985). *Measurement and evaluation in teaching* (6th ed.). New York, NY: Macmillan Publishing Company.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York, NY: McGraw-Hill.

Guion, R. M. (1965). *Personnel testing*. New York, NY: McGraw-Hill.

Gullicksen, Harold. (1950). *Theory of mental tests*. New York, NY: John Wiley & Sons.

Hambleton, Ronald K. (Ed.). (1983). *Application of item response theory*. Vancouver, British Columbia, Canada: Educational Research Institute of British Columbia.

Hambleton, Ronald K. (1989). Principles and selected applications of item response theory. In Robert L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 147-200). New York, NY: American Council on Education; & Macmillan Publishing Company.

Hambleton, Ronald K.; & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

Hambleton, Ronald K.; Swaminathan, H.; & Rogers, H. Jane. (1991). *Fundamentals of item response theory* (Measurement Methods for the Social Sciences Series, Vol. 2). Newbury Park, CA: Sage Publications, Inc.

Hambleton, Ronald K.; & Zaal, J. (Eds.). (1988). *Handbook of testing*. Amsterdam, The Netherlands: North Holland Press.

Harrow, A. J. (1972). *A taxonomy of the psychomotor domain: A guide for developing behavioral objectives*. New York, NY: David McKay.

Harvard Business Review. (1991). *Appraising performance appraisal* (Harvard Business Review Paperback No. 90070). Boston, MA: Harvard Business School Press.

Haynes, S. N.; & Wilson, C. C. (1979). *Behavioral assessment*. San Francisco, CA: Jossey-Bass, Inc., Publishers.

Heneman, Herbert G., III; Schwab, Donald P.; Fossum, J. A.; & Dyer, L. D. (1989). *Personnel/human resources management* (4th ed.). Homewood, IL: Richard D. Irwin, Inc.

Herman, Joan L.; Aschbacher, Pamela R.; & Winters, Lynn. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Hogan, Joyce; & Hogan, Robert. (1990). *Business and industry testing: Current practices and test reviews*. Austin, TX: PRO-ED, Inc.

Holland, Paul W.; & Rubin, D. R. (Eds.). (1982). *Test equating*. New York, NY: Academic Press.

Hopkins, K. D.; & Stanley, J. C. (1981). *Educational and psychological measurement and evaluation* (6th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Hulin, C. L.; Drasgow, F.; & Parsons, C. K. (1983). *Item response theory: Applications to psychological measurement*. Homewood, IL: Dow Jones-Irwin.

Jacobs, Lucy Cheser; & Chase, Clinton I. (1992). *Developing and using tests effectively: A guide for faculty*. San Francisco, CA: Jossey-Bass, Inc., Publishers.

Jensen, Arthur R. (1980). *Bias in mental testing*. New York, NY: Free Press.

Kamii, Constance. (Ed.). (1989). *Achievement testing in the early grades*. Washington, DC: National Association for the Education of Young Children.

Krathwohl, D. R.; Bloom, Benjamin S.; & Masia, B. B. (1964). *Taxonomy of educational objectives: The classification of educational goals: Handbook 2: Affective domain*. New York, NY: David McKay.

Lake, Dale G.; Miles, Matthew B.; & Earle, Ralph B. (1973). *Measuring human behavior*. New York, NY: Columbia University, Teachers College Press.

Landy, Frank J.; & Farr, James L. (1983). *The measurement of work performance: Methods, theory, and applications*. New York, NY: Academic Press.

Latham, G. P.; & Wexley, K. N. (1981). *Increasing productivity through performance appraisal.* Reading, MA: Addison-Wesley.

Lindquist, E. F. (Ed.). (1951). *Educational measurement.* Washington, DC: American Council on Education.

Linn, Robert L. (Ed.). (1989). *Educational measurement* (3rd ed.). New York, NY: American Council on Education; & Macmillan Publishing Company.

Livingston, Samuel A.; & Zieky, Michael J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

Lord, Frederic M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

Lord, Frederic M.; & Novick, Melvin R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley Publishing Company.

Mehrens, William A.; & Lehmann, Irvin J. (1987). *Using standardized tests in education* (4th ed.). New York, NY: Longman.

Mehrens, William A.; & Lehmann, Irvin J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). Troy, MO: Holt, Rinehart & Winston.

Messick, Samuel. (1989). Validity. In Robert L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York, NY: American Council on Education; & Macmillan Publishing Company.

Millman, Jason; & Darling-Hammond, Linda. (Eds.). (1990). *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers*. Newbury Park, CA: Sage Publications, Inc.

Mitchell, James V., Jr.; Wise, Steven L.; & Plake, Barbara S. (Eds.). (1990). *Assessment of teaching: Purposes, practices, and implications for the profession.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

Morris, Lynn Lyons; Fitz-Gibbons, Carol Taylor; & Lindheim, Elaine. (1987). *How to measure performance and use tests.* Newbury Park, CA: Sage Publications, Inc.

National Commission on Testing and Public Policy. (1989). *From gatekeeper to gateway: Transforming testing in America.* Chestnut Hill, MA: Author, Boston College.

Nitko, Anthony J. (1983). *Educational tests and measurement: An introduction.* New York, NY: Harcourt Brace Jovanovich, Inc.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Office of Technology Assessment (Congress of the United States). (1992). *Testing in American schools: Asking the right questions* (Summary Report). Washington, DC: Author.

Osterlind, Steven J. (1989). *Constructing test items.* Boston, MA: Kluwer Academic.

Petersen, Nancy S.; Kolen, Michael J.; & Hoover, H. D. (1989). Scaling, norming, and equating. In Robert L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 221-262). New York, NY: American Council on Education; & Macmillan Publishing Company.

Plake, Barbara S. (Ed.). (1984). *Social and technical issues in testing: Implications for test construction and usage*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

Plake, Barbara S.; & Witt, Joseph. (Eds.). (1986). *The future of testing*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

Popham, W. James. (1978). *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Priestley, M. (1982). *Performance assessment in education and training: Alternative techniques*. Englewood Cliffs, NJ: Educational Technology Publications.

Rasch, Georg. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: The University of Chicago Press.

Reynolds, Cecil R.; & Kamphaus, Randy W. (Eds.). (1990). *Handbook of psychological and educational assessment of young children*. New York, NY: Guilford Press.

Roid, Gale H.; & Haladyna, Thomas M. (1982). *A technology of test-item writing*. Orlando, FL: Academic Press.

Schmitt, Neal; Borman, Walter C.; & Associates. (1992). *Personnel selection in organizations* (Frontiers of Industrial and Organizational Psychology Series No. 6). San Francisco, CA: Jossey-Bass, Inc., Publishers.

Shavelson, Richard J.; & Webb, Noreen M. (1991). *Generalizability theory: A primer* (Measurement Methods for the Social Sciences Series, Vol. 1). Newbury Park, CA: Sage Publications, Inc.

Shaw, M. E.; & Wright, J. M. (1967). *Scales for the measurement of attitudes*. New York, NY: McGraw-Hill.

Shepard, Lorrie A. (1993). Evaluating test validity. In Linda Darling-Hammond (Ed.), *Review of research in education* (Vol. 19) (pp. 405-450). Washington, DC: American Educational Research Association.

Shimberg, Benjamin. (1982). *Occupational licensing: A public perspective*. Princeton, NJ: Educational Testing Service.

Society for Industrial and Organizational Psychology, Inc. (1987). *Principles for the validation and use of personnel selection procedures* (3rd ed.). College Park, MD: Author.

Spandel, Vicki; & Stiggins, Richard J. (1990). *Creating writers: Linking assessment and writing instruction*. New York, NY: Longman.

Stronge, James H.; & Helm, Virginia M. (1991). *Evaluating professional support personnel in education*. Newbury Park, CA: Sage Publications, Inc.

Thorndike, Robert L. (1949). *Personnel selection test and measurement techniques*. New York, NY: Wiley.

Thorndike, Robert L. (Ed.). (1971). *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Thorndike, Robert L. (1982). *Applied psychometrics*. Boston, MA: Houghton-Mifflin.

Thornton, George C., III; & Byham, William C. (Eds.). (1982). *Assessment centers and managerial performance*. New York, NY: Academic Press.

Tyler, Leona E. (1963). *Tests and measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Wainer, Howard; & Braun, Henry I. (Eds.). (1988). *Test validity*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

Wainer, Howard; & Messick, Samuel. (Eds.). (1983). *Principles of modern psychological measurement: A Festschrift for Frederic M. Lord*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

Walberg, Herbert J.; & Haertel, Geneva D. (Eds.). (1991). *The international encyclopedia of educational evaluation*. Oxford, England: Pergamon Press.

Weiss, D. J. (Ed.). (1983). *New horizons in testing: Latent trait theory and computerized adaptive testing*. New York, NY: Academic Press.

Wittrock, Merlin C.; & Baker, Eva L. (Eds.). (1991). *Testing and cognition*. Englewood Cliffs, NJ: Prentice Hall.

Wright, Benjamin D.; & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago, IL: MESA Press.

Articles:

American Federation of Teachers; National Council on Measurement in Education; & National Education Association. (1990, Winter). Standards for teacher competence in educational assessment of students. *Educational Measurement: Issues and Practice*, *9*(4), 30-32.

Anastasi, Anne. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, *37*, 1-15.

Anderson, Richard C. (1972). How to construct achievement tests to assess comprehension. *Review of Educational Research*, *42*, 145-170.

Angoff, William H. (1987). Technical and practical issues in equating: A discussion of four papers. *Applied Psychological Measurement*, *11*, 291-300.

Angoff, William H.; & Cowell, William R. (1986). An examination of the assumption that the equating of parallel forms is population-independent. *Journal of Educational Measurement*, *23*, 327-345.

Banta, Trudy W. (1988). Implementing outcomes assessment: Promises and perils. *New Directions for Institutional Research*, *59*.

Banta, Trudy W.; & Fisher, H. S. (1987, March 4). Measuring how much students have learned entails much more than simply testing them. *Chronicle of Higher Education*, p. 45.

Barrs, Myra. (1990). The primary language record: Reflections of issues in evaluation. *Language Arts*, *67*(3), 244-253.

Berk, Ronald A. (1986, Spring). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56*(1), 137-172.

Bock, R. Darrell; Mislevy, Robert; & Woodson, Charles. (1982, March). The next stage in educational assessment. *Educational Researcher, 11*(3), 4-11.

Carter, K. (1986). Test-wiseness for teachers and students. *Educational Measurement: Issues and Practice, 5*, 20-23.

Cole, Nancy S. (1981). Bias in testing. *American Psychologist, 36*(10), 1067-1077.

Cronbach, Lee J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Ebel, Robert L. (1951). Estimation of the reliability of ratings. *Psychometrika, 16*, 297-334.

Feldt, Leonard S.; Steffen, M.; & Gupta, N. C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement, 9*, 351-361.

Flanagan, John C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327-358.

Fremer, John; Diamond, Esther; & Camara, Wayne. (1989). Developing a code of fair testing practices in education. *American Psychologist, 44*(7), 1062-1067.

Glass, Gene V. (1978). Standards and criteria. *Journal of Educational Measurement, 15*(4), 237-261.

Guttman, L. A. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*, 255-282.

Haertel, Edward. (1985). Construct validity and criterion-referenced testing. *Review of Educational Research, 55*, 23-46.

Hakel, M. D. (1986). Personnel selection and placement. *Annual Review of Psychology, 37*, 351-380.

Haladyna, Thomas M.; Nolan, Susan Bobbitt; & Haas, Nancy S. (1991, June/July). Raising standardized achievement test scores and the origins of test pollution. *Educational Researcher, 20*(5), 2-8.

Hills, John R. (1989, Winter). Screening for potentially biased items in testing programs. *Educational Measurement: Issues and Practice, 8*(4), 5-11.

Hunter, J. E.; & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72-98.

Kane, Michael T. (1982). A sampling model for validity. *Applied Psychological Measurement, 6*, 125-160.

Kuder, G. F.; & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151-160.

Landy, Frank J.; & Farr, James L. (1980). Performance ratings. *Psychological Bulletin, 87*, 72-107.

Linn, Robert L. (1993, Spring). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*(6), 1-16.

Linn, Robert L.; Baker, Eva L.; & Dunbar, Stephen B. (1991, November). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Lord, Frederic M. (1955). Estimating test reliability. *Educational and Psychological Measurement, 15*, 325-336.

Lord, Frederic M. (1963). Cutting scores and errors of measurement. *Psychometrika, 27*(1), 19-30.

Lord, Frederic M. (1965). A strong true score theory with applications. *Psychometrika, 30*, 59-72.

Lord, Frederic M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14*, 117-138.

Lord, Frederic M. (1984). Standard errors of measurement at different score levels. *Journal of Educational Measurement, 21*, 239-243.

McEntee, G. H. (1990, April). When the teacher has to choose: What if you're told to give a test you know is bad? *Teacher Magazine*, 8-9.

Mehrens, William A. (1992, Spring). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11*(1), 3-9, 20.

Mehrens, William A.; & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent? *Educational Measurement: Issues and Practice, 8*, 14-22.

Messick, Samuel. (1981, November). Evidence and ethics in the evaluation of tests. *Educational Researcher*, *10*(9), 9-20.

Millman, Jason. (1989, August/September). If at first you don't succeed: Setting passing scores when more than one attempt is permitted. *Educational Researcher*, *18*(6), 5-9.

Millman, Jason; Bishop, C. H.; & Ebel, Robert L. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement*, *25*, 707-726.

Mills, Craig N. (1983). A comparison of three methods of establishing cut-off scores on criterion-referenced tests. *Journal of Educational Measurement*, *20*, 283-292.

Mills, R. P. (1989, December). Portfolios capture rich array of student performance. *School Administrator*, *11*, 8-11.

Moss, Pamela A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, *62*, 229-258.

Novick, Melvin R.; & Ellis, D. D. (1977). Equal opportunity in educational and employment selection. *American Psychologist*, *32*, 306-320.

Paulson, F. Leon; Paulson, Pearl; & Meyer, Carol. (1991). What makes a portfolio a portfolio? *Educational Leadership*, *48*(5), 60-63.

Phillips, Gary W. (1990, Fall). The Lake Wobegon effect. *Educational Measurement: Issues and Practice*, *9*(3), 3, 14

Popham, W. James. (1987). The merits of measurement-driven instruction. *Phi Delta Kappan*, *68*(9), 679-682.

Reynolds, C. R. (1984-1985). Critical measurement issues in learning disabilities. *Journal of Special Education, 18*, 451-476.

Schmidt, F. L.; & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. *American Psychologist, 36*, 1128-1137.

Schwab, Donald P.; Heneman, Herbert G., III; & DeCotis, Thomas A. (1975). Behaviorally anchored rating scales: A review of the literature. *Personnel Psychology, 28*, 549-562.

Scriven, Michael. (1987). Validity in personnel evaluation. *Journal of Personnel Evaluation in Education, 1*(1), 9-23.

Shavelson, Richard J.; Baxter, Gail P.; & Pine, Jerry. (1992, May). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22-27.

Shavelson, Richard J.; Webb, Noreen M.; & Rowley, Glenn L. (1989). Generalizability theory. *American Psychologist, 44*, 922-932.

Shepard, Lorrie A. (1980). Standard setting issues and methods. *Applied Psychological Measurement, 4*, 447-467.

Shepard, Lorrie A. (1989, April). Why we need better assessments. *Educational Leadership, 46*(7), 4-9.

Skaggs, Gary; & Lissitz, Robert W. (1987). IRT test equating: Relevant issues and a review of recent literature. *Review of Educational Research, 56*, 495-529.

Stiggins, Richard J. (1988, January). Revitalizing classroom assessment: The instructional priority. *Phi Delta Kappan, 69*(5), 363-368.

Westbrook, Bert W.; & Mastie, Marjorie M. (1983, Spring). Doing your homework: Suggestions for the evaluation of tests by practitioners. *Educational Measurement: Issues and Practice*, *2*(1), 11-14, 26.

Wiggins, Grant. (1989, April). Teaching to the (authentic) test. *Educational Leadership*, *46*(7), 41-47.

Wiggins, Grant. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*, 703-713.

Wolf, Dennie R. (1989, April). Portfolio assessment: Sampling student work. *Educational Leadership*, *46*(7), 35-36.

Wright, Benjamin D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*, 97-116.

Wright, Benjamin D.; & Panchapakesan, W. A. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, *29*, 23-48.

Yen, Wendy M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, *23*, 299-325.

NCME Instructional Modules:

Since 1987, the National Council on Measurement in Education (NCME) has provided a series of instructional units, called ITEMS, in its quarterly publication, *Educational Measurement: Issues and Practice.* The list of ITEMS through the fall of 1993 is provided below.

Arter, Judith A.; & Spandel, Vicki. (1992, Spring). Using portfolios of student work in instruction and assessment. *11*(1), 36-44.

Brennan, Robert L. (1992, Winter). Generalizability theory. *11*(4), 27-34.

Brookhart, Susan. (1993, Spring). Assessing student achievement with term papers and written reports. *12*(1), 40-47.

Cook, Linda L.; & Eignor, Daniel R. (1991, Fall). IRT equating models. *10*(3), 37-45.

Frary, Robert R. (1988, Summer). Formula scoring of multiple choice tests (Correction for guessing). *7*(2), 33-38.

Frisbie, David A. (1988, Spring). Reliability of scores from teacher-made tests. *7*(1), 25-35.

Frisbie, David A.; & Waltman, Kristie K. (1992, Fall). Developing a personal grading plan. *11*(3), 35-42.

Hambleton, Ronald K.; & Jones, Russell W. (1993, Fall). Comparison of classical test theory and item response theory. *12*(3), 38-47.

Harris, Debrah. (1989, Spring). Comparison of 1-, 2-, and 3-parameter IRT models. *8*(1), 35-41.

Harvill, Leo M. (1991, Summer). Standard error of measurement. *10*(2), 33-41.

Kolen, Michael J. (1988, Winter). Traditional equating methodology. *7*(4), 29-36.

Oosterhof, Albert C. (1987, Winter). Obtaining intended weights when combining student scores. *6*(4), 29-37.

Osterlind, Steven. (1988, Fall). Using CRTs in program curriculum evaluation. *7*(3), 23-30.

Stiggins, Richard J. (1987, Fall). Design and development of performance assessments. *6*(3), 33-42.

Stiggins, Richard J. (1992, Summer). High quality classroom assessment: What does it really mean? *11*(2), 35-39.

Traub, Ross E.; & Rowley, Glenn L. (1991, Spring). Understanding reliability. *10*(1), 37-45.

Sources of Information on Tests:

Chun, K. T.; Cobb, S.; & French, J. R. P. (1975). *Measures for psychological assessment: A guide to 3,000 original sources and their applications.* Ann Arbor, MI: University of Michigan, Institute for Social Research.

Cook, John D.; Hepworth, Susan J.; Wall, Toby D.; & Warr, Peter B. (1981). *The experience of work: A compendium and review of 249 measures and their use.* New York, NY: Academic Press.

Educational Testing Service. (1989-1992). *The ETS Test Collection catalog*. Phoenix AZ: Oryx Press.
Volume 1: Achievement Tests (2nd ed.) (1992)
Volume 3: Tests for Special Populations (1989)
Volume 4: Cognitive Aptitude Tests (1990)
Volume 5: Attitude Tests (1991)
Volume 6: Personality Tests (1992)
Volume 2: Vocational Measures is out of print.

Goldman, Bert A.; & Mitchell, David F. (1990). *Directory of unpublished experimental mental measures*. Dubuque, IA: William C. Brown, Publishers.

Kapes, Jerome T.; & Mastie, Marjorie Moran. (1988). *A counselor's guide to career assessment instruments* (2nd ed.). Alexandria, VA: National Career Development Association.

Keyser, Daniel J. & Sweetland, Richard C. (Eds.). (1984-92). *Test* (Volumes I - IX). Austin, TX: PRO-ED.

Kramer, Jack J.; & Conoley, Jane Close. (Eds.). (1992). *The eleventh mental measurements yearbook*. Lincoln, NE: The University of Nebraska--Lincoln, Buros Institute of Mental Measurements.

Krug, Samuel E. (Ed.). (1988-1989). *Psychware sourcebook* (3rd ed.). Austin, TX: PRO-ED., Inc.

Mitchell, James V., Jr. (1983). *Tests in print* (3rd ed.). Lincoln, NE: The University of Nebraska--Lincoln, Buros Institute for Mental Measurements. (4th edition to be published in 1994)

Science Directorate of the American Psychological Association. (1993). *Finding information about psychological tests: A guide for locating and using both published and unpublished tests*. Washington, DC: Author.

Stoloff, Michael L.; & Couch, James V. (1992). *Computer use in psychology: A directory of software* (3rd ed.). Washington, DC: American Psychological Association.

Sweetland, Richard C.; & Keyser, Daniel J. (Eds.). (1991). *Tests: A comprehensive reference for assessments in psychology, education, and business* (3rd ed.). Austin, TX: PRO-ED, Inc.

The ETS Test Collection has compiled over 200 annotated bibliographies of various types of tests. They cover several major categories of tests including achievement, aptitude, attitudes and interests, personality, sensory-motor, special populations, vocational/occupational, and miscellaneous tests. Write to: Test Collection, Educational Testing Service, Princeton, NJ 08541-0001.

Other Reference Books:

Filkins, James H.; & Caruth, Donald L. (1973). *Lexicon of American business terms*. New York, NY: Simon and Schuster, Inc.

Scriven, Michael. (1991). *The evaluation thesaurus* (4th ed.). Newbury Park, CA: Sage Publications, Inc.

Vogt, W. Paul. (1993). *Dictionary of statistics and methodology*. Newbury Park, CA: Sage Publications, Inc.

Wheeler, Patricia; Haertel, Geneva D.; & Scriven, Michael. (1992). *Teacher evaluation glossary*. Kalamazoo, MI: Western Michigan University, The Evaluation Center.

Yaremko, R. M.; Harari, Herbert; Harrison, Robert C.; & Lynn, Elizabeth. (1986). *Handbook of research and quantitative methods in psychology: For students and professionals*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

STANDARDS FOR ASSESSMENT PROGRAMS AND TESTS

Below we have listed several sets of standards and provided an overview of what these various standards cover. The document with the full set of standards and explanatory materials is listed. Assessment program developers and users should be familiar with the full documents.

These are not the only sets of standards available. Testing programs, licensing and certification boards, and professional associations may have standards for their own assessments programs and tests. For example, the National Board for Professional Teaching Standards has five categories of guidelines for the development of its certification assessments: administratively feasible, professionally credible, publicly acceptable, legally defensible, and economically affordable.

American Educational Research Association; American Psychological Association; & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

This is the fifth set of guidelines developed by the three sponsoring organizations. They have been written by The Committee to Develop Standards for Educational and Psychological Testing, and fall into four groups of standards, each group with two or more categories, and each category with several standards in it. The groups and their categories, along with the number of standards in each category, are as follows:

Technical Standards
Validity (25)
Reliability and Errors of Measurement (12)
Test Development and Revision (25)
Scaling, Norming, Score Comparability, & Equating (9)
Test Publication: Technical Manuals, User's Guides (11)

Professional Standards for Test Use
General Principles of Test Use (13)
Clinical Testing (6)
Educational, Psychological Testing in Schools (12)
Test Use in Counseling (9)
Employment Testing (9)
Professional, Occupational Licensure & Certification (5)
Program Evaluation (8)

Standards for Particular Applications
Testing Linguistic Minorities (7)
Testing People Who Have Handicapping Conditions (8)

Standards for Administrative Procedures
Test Administration, Scoring, and Reporting (11)
Protecting the Rights of Test Takers (10)

Educational Testing Service. (1987). *ETS standards for quality and fairness*. Princeton, NJ: Author.

These standards were developed by ETS as corporate policy in response to the AERA/APA/NCME standards (described above) and were adopted by the ETS Board of Trustees. The standards are reflected by a statement of principle, a list of policies, and a set of procedural guidelines for all ETS tests and testing programs. There are seven categories, listed below:

Accountability
Confidentiality of Data
Quality Control for Accuracy and Timeliness
Tests and Measurement--Technical Quality of Tests:
- Validity
- Test Development
- Test Administration
- Reliability
- Scale Definition
- Equating
- Score Interpretation

Test Use
Research and Development
Public Information

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor and the Department of Justice. (1978, August 25). Uniform guidelines on employee selection procedures. *Federal Register*, *43*(166), 38290-38315.

These Guidelines have been issued jointly by four federal agencies and are the basis for litigation with respect to personnel assessment. The first part of the Guidelines covers the purpose and scope of the Guidelines, defines discrimination and adverse impact, and sets the standards for validity for employee selection devices. Next they cover the relationship between validity studies and information about the job, criterion-related validity, content validity, and construct validity. (It should be noted that these definitions of validity are not necessarily consistent with the AERA/APA/NCME standards listed above.) The Guidelines also address documentation and impact of validity evidence. Although the Guidelines are not law, they are treated as if they were law, based on the U.S. Supreme Court's rulings in *Griggs v. Duke Power Co.* (1971) and *Albemarle Paper Co. v. Moody* (1975) that the Guidelines are entitled to "great deference."

Joint Committee on Standards for Educational Evaluation. (1988). *The personnel evaluation standards: How to assess systems for evaluating educators.* Newbury Park, CA: Sage Publications, Inc.

This document provides 21 standards in four areas for systems for evaluating educators.

Propriety
Service Orientation
Formal Evaluation Guidelines
Conflict of Interest
Access to Personnel Evaluation Reports
Interactions with Evaluatees

Utility
Constructive Orientation
Defined Uses
Evaluator Credibility
Functional Reporting
Follow-up and Impact

Feasibility
Practical Procedures
Political Viability
Fiscal Viability

Accuracy
Defined Role
Work Environment
Documentation of Procedures
Valid Measurement
Reliable Measurement
Systematic Data Control
Bias Control
Monitoring Evaluation Systems

Joint Committee on Testing Practices. (1988). *Code of fair testing practices in education.* Washington, DC: Author, American Psychological Association.

This Code, designed to cover only educational testing, is concerned with the roles of the test developer and the test user. For each of the four categories, standards are provided for both developers and users. The categories and numbers of standards are as follows:

Developing/Selecting Appropriate Tests (8 for each)
Interpreting Scores (5 for each)
Striving for Fairness (3 for each)
Informing Test Takers (2 for developers, 3 for users)

Society for Industrial and Organizational Psychology, Inc. (1987). *Principles for the validation and use of personnel selection procedures* (3rd ed.). College Park, MD: Author.

This document brings together earlier editions of the *Principles* and the 1985 *Standards for Educational and Psychological Testing* of AERA/APA/NCME (first one listed in this section) as well as recent research and theoretical developments. Several topics are covered by the principles. They are listed below:

Objectives of the Validation Efforts
Job Analysis
Criterion-Related Strategy
Differential Prediction
Content-Oriented Strategies
Construct-Oriented Strategies
Validity Generalization
Operational Use

REFERENCES

The publications listed below were used in preparing this *Resource Handbook*.

American Educational Research Association; American Psychological Association; & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Anastasi, Anne. (1954). *Psychological testing*. New York, NY: The Macmillan Company.

Angoff, William H.; & Dyer, Henry S. (1971). The Admissions Testing Program. In William H. Angoff (Ed.), *The College Board Admissions Testing Program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Tests* (pp. 1-13). New York, NY: College Entrance Examination Board.

Berk, Ronald A. (Ed.). (1986). *Performance assessment: Methods and applications*. Baltimore, MD: The Johns Hopkins University Press.

Boring, Edwin G. (1957). *A history of experimental psychology* (2nd ed.). New York, NY: Appleton-Century-Crofts, Inc.

Butts, R. Freeman; & Cremin, Lawrence A. (1953). *A history of education in American culture*. New York, NY: Holt, Rinehart & Winston.

California Test Bureau. (undated). *A glossary of measurement terms*. Monterey, CA: Author.

Cannell, John J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all fifty states are above the national average*. Daniels, WV: Friends for Education.

Chauncey, Henry; & Dobbin, John E. (1963). *Testing: Its place in education today*. New York, NY: Harper & Row, Publishers.

Crocker, Linda; & Algina, James. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Holt, Rinehart & Winston.

Cronbach, Lee J. (1960). *Essentials of psychological testing* (2nd ed.). New York, NY: Harper & Brothers, Publishers.

DuBois, Philip H. (1966). A test-dominated society: China, 1115 B.C.-1905 A.D. In Anne Anastasi (Ed.), *Testing problems in perspective: Twenty-fifth anniversary volume of topical readings from the Invitational Conference on Testing Problems*. Washington, DC: American Council on Education.

Educational Testing Service. (1987). *ETS standards for quality and fairness*. Princeton, NJ: Author.

English, Horace B.; & English, Ava Champney. (1958). *A comprehensive dictionary of psychological and psychoanalytical terms: A guide to usage*. New York, NY: Longmans, Green and Company.

Filkins, James H.; & Caruth, Donald L. (1973). *Lexicon of American business terms*. New York, NY: Simon and Schuster, Inc.

Gardner, Howard. (1985). *Frames of mind: The theory of multiple intelligences* (2nd ed.). New York, NY: Basic Books.

Haladyna, Thomas M.; Nolen, Susan Bobbit; & Hass, Nancy S. (1991, June/July). Raising standardized achievement test scores and the origin of score pollution. *Educational Researcher, 20*(5), 2-8.

Hambleton, Ronald K.; Swaminathan, H.; & Rogers, H. Jane. (1991). *Fundamentals of item response theory* (Measurement Methods for the Social Science Series, Vol. 2). Newbury Park, CA: Sage Publications, Inc.

Hawkins, Joyce M.; & Allen, Robert. (Eds.). (1991). *The Oxford encyclopedic English dictionary*. New York, NY: Oxford University Press.

Jacobs, Lucy Cheser; & Chase, Clinton I. (1992). *Developing and using tests effectively: A guide for faculty*. San Francisco, CA: Jossey-Bass, Inc., Publishers.

Jaeger, Richard M. (1990). *Statistics: A spectator sport* (2nd ed.). Newbury Park, CA: Sage Publications, Inc.

Joint Committee on Standards for Educational Evaluation. (1988). *The personnel evaluation standards: How to assess systems for evaluating educators*. Newbury Park, CA: Sage Publications, Inc.

Kramer, Jack J.; & Conoley, Jane Close. (Eds.). (1992). *The eleventh mental measurements yearbook*. Lincoln, NE: The University of Nebraska--Lincoln, Buros Institute of Mental Measurements.

Linn, Robert L. (Ed.). (1989). *Educational measurement* (3rd ed.). New York, NY: American Council on Education; & Macmillan Publishing Company.

Linn, Robert L. (1993, Spring). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*(6), 1-16.

Linn, Robert L.; Baker, Eva L.; & Dunbar, Stephen B. (1991). Complex, performance-based assessment: Expectation and validation criteria. *Educational Researcher, 20*(8), 15-21.

Madaus, George F.; Haney, Walt; & Kreitzer, Amelia. (1992). *Testing and evaluation: Learning from the projects we fund*. New York, NY: Council for Aid to Education.

McNemar, Quinn. (1962). *Psychological statistics* (3rd ed.). New York, NY: John Wiley and Sons, Inc.

Mehrens, William A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11*(1), 3-9, 20.

Messick, Samuel. (1989). Validity. In Robert L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York, NY: American Council on Education; & Macmillan Publishing Company.

Mitchell, Blythe C. (undated). *A glossary of measurement terms* (Test Service Notebook 13). New York, NY: Harcourt Brace Jovanovich, Inc., Test Department.

Mitchell, Ruth. (1992). *Testing for learning: How new approaches to evaluation can improve American schools*. New York, NY: The Free Press, A Division of Macmillan, Inc.

Morris, Lynn Lyons; Fitz-Gibbons, Carol Taylor; & Lindheim, Elaine. (1987). *How to measure performance and use tests*. Newbury Park, CA: Sage Publications Inc.

New Webster's dictionary and thesaurus of the English language. (1992). Danbury, CT: Lexicon Publications, Inc.

Nigro, Felix A.; & Nigro, Lloyd G. (1976). *The new public personnel administration*. Itasca, IL: F. E. Peacock Publishers, Inc.

Office of Technology Assessment (Congress of the United States). (1992). *Testing in American schools: Asking the right questions* (Summary Report). Washington, DC: Author.

Rudman, Herbert C. (1987, Fall). The future of testing is now. *Educational Measurement: Issues and Practice*, *6*(3), 5-11.

Schubert, William H. (1986). *Curriculum: Perspective, paradigm, and possibilities*. New York, NY: Macmillan Publishing Company.

Science Directorate of the American Psychological Association. (1993). *Finding information about psychological tests: A guide for locating and using both published and unpublished tests*. Washington, DC: Author.

Scriven, Michael. (1991). *The evaluation thesaurus* (4th ed.). Newbury Park, CA: Sage Publications, Inc.

Shavelson, Richard J.; Baxter, Gail P.; & Pine, Jerry. (1992, May). Performance assessment: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22-27.

Shimberg, Benjamin. (1982). *Occupational licensing: A public perspective.* Princeton, NJ: Educational Testing Service, Center for Occupational and Professional Assessment.

Society for Industrial and Organizational Psychology, Inc. (1987). *Principles for the validation and use of personnel selection procedures* (3rd ed.). College Park, MD: Author.

Super, Donald E. (1949). *Appraising vocational fitness by means of psychological tests.* New York, NY: Harper & Brothers.

Test Collection. (1992, January). *Major U.S. publishers of standardized tests.* Princeton, NJ: Educational Testing Service.

Vogt, W. Paul. (1993). *Dictionary of statistics and methodology.* Newbury Park, CA: Sage Publications, Inc.

Walberg, Herbert J. (1990). National Assessment for Educational Progress: Retrospect and prospect. In Herbert J. Walberg & Geneva D. Haertel (Eds.), *The international encyclopedia of educational evaluation* (pp. 440-443). Oxford, England: Pergamon Press.

Walberg, Herbert J.; & Haertel, Geneva D. (Eds.). (1991). *The international encyclopedia of educational evaluation.* Oxford, England: Pergamon Press.

Walsh, Michael F.; Loret, Peter G.; et al. (1979, December). *Handbook for proficiency assessment* (1st ed.). Berkeley, CA: Educational Testing Service. (Developed for the California State Department of Education)

Webster's new collegiate dictionary. (1956). Springfield, MA: G. & C. Merriam Co., Publishers.

Wheeler, Patricia. (1990). *Normal curve equivalents (NCEs)* (Tip Sheet). Mountain View, CA: RMC Research Corporation, Chapter 1 Region F TAC.

Wheeler, Patricia; Haertel, Geneva D.; & Scriven, Michael. (1992). *Teacher evaluation glossary.* Kalamazoo, MI: Western Michigan University, The Evaluation Center.

Wittrock, Merlin C.; & Baker, Eva L. (Eds.). (1991). *Testing and cognition.* Englewood Cliffs, NJ: Prentice Hall.

Yaremko, R. M.; Harari, Herbert; Harrison, Robert C.; & Lynn, Elizabeth. (1986). *Handbook of research and quantitative methods in psychology: For students and professionals.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

213

ABOUT THE AUTHORS

Patricia Wheeler is the President of EREAPA Associates in Livermore, California and a Senior Research Associate for the Center for Research on Educational Accountability and Teacher Evaluation (CREATE) at Western Michigan University in Kalamazoo. She was a researcher, evaluator, technical consultant, and testing program administrator for almost 24 years at Educational Testing Service. She has a Ph.D. in Policy Analysis from the Graduate School of Education at the University of California at Berkeley. In addition, she has an Ed.M. in Counseling, an M.B.A. in Management, and an A.B. in Psychology. She has written over 50 reports, papers, and journal articles. She belongs to several professional associations including the American Educational Research Association, the American Evaluation Association, the National Association of Secondary School Principals, and the National Council on Measurement in Education.

Geneva D. Haertel is a Research Associate for Temple University's Center for Research on Human Development and Education and for Western Michigan University's Center for Research on Educational Accountability and Teacher Evaluation (CREATE). Also, she is an independent consultant for EREAPA Associates. She has worked at the R & D Center for Cognitive Learning at the University of Wisconsin--Madison, for the research and evaluation unit of the Chicago Board of Education, and as a research associate at the University of Illinois at Chicago. She has developed tests and evaluated curricula and

instructional programs. Her primary area of expertise is the educational achievement of regular and disadvantaged students. She has a Ph.D. in educational psychology from Kent State University. Dr. Haertel has published over 35 journal articles and co-edited *The International Encyclopedia of Educational Evaluation* with Herbert Walberg. She is a member of the American Educational Research Association, the American Evaluation Association, and the Association for Supervision and Curriculum Development.

Order additional copies at the low first-printing price.

Resource Handbook on Performance Assessment *by Wheeler & Haertel*

Wheeler & Haertel, evaluators and academicians, have brought together an impressive handbook that helps to demystify the confusing terms used in measurement and assessment. This 232-page volume includes a glossary of over 500 cross-referenced terms, a historical overview, and a rich resources section providing names of associations and agencies, test publishers, references, directories of test information, and instructional modules. *Professionals in business, education, government, health, and the military will find this volume to be a valuable guide.*

Order ISBN 1-884690-13-7. The Owl Press, P.O. Box 89, Berkeley, CA 94701 (FAX-486-0743) Price is $18.95 (Calif. residents please include 8.25% sales tax) prepaid. Discounts available on orders of 10 or more. Please provide shipping information:

Number of Books:

Prepaid $18.95 ea. plus tax:

Name

Organization

Shipping Address

City, State, Zip

Phone or FAX

Amount enclosed:

Resource Handbook on Performance Assessment and Measurement:
A Tool for Students, Practitioners, and Policymakers

by Patricia Wheeler and Geneva D. Haertel

Reviewers Comments:

"An invaluable guide for new teachers becoming oriented to schools and assessment practices." Rasjidah Franklin, U.C. Berkeley

"This handbook clarifies the meaning of many commonly used, but often misunderstood terms. It should be of great value to teachers, school administrators, and others." Jerry Jenkins, Educational Testing Service

"Students and professionals will find this comprehensive glossary very helpful--definitive!" Gale Roid, George Fox College and Oregon Department of Education

---

The Owl Press
P.O. Box 89
Berkeley, CA 94701

BEST COPY AVAILABLE